2023

EXPLICIT RULE LEARNING: A COGNITIVE TUTORIAL METHOD TO TRAIN USERS OF ARTIFICIAL INTELLIGENCE/MACHINE LEARNING SYSTEMS

Anne Linja
Michigan Technological University, alinja@mtu.edu

Recommended Citation

Linja, Anne, "EXPLICIT RULE LEARNING: A COGNITIVE TUTORIAL METHOD TO TRAIN USERS OF ARTIFICIAL INTELLIGENCE/MACHINE LEARNING SYSTEMS", Open Access Dissertation, Michigan Technological University, 2023.
https://doi.org/10.37099/mtu.dc.etdr/1642

EXPLICIT RULE LEARNING: A COGNITIVE TUTORIAL METHOD TO TRAIN USERS OF ARTIFICIAL INTELLIGENCE/MACHINE LEARNING SYSTEMS

By

Anne Linja

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Dissertation Advisor:	*Shane T. Mueller*
Committee Member:	*Elizabeth S. Veinott*
Committee Member:	*Laura E. Brown*
Committee Member:	*Leo C. Ureel II*
Department Chair:	*Kelly S. Steelman*

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to thank Dr. Shane Mueller for the privilege of being his advisee. I have learned so much from his classes, but that was just the beginning. He inspires me to go beyond hard work, and patiently guides me to research and learn things on my own, helping me to achieve a higher understanding than I would have otherwise. There's so much more to learn, but every day I learn how amazingly awesome cognitive science is, and I thank Dr. Mueller for helping me discover and delve deeper into this domain.

Thank you, Dr. Elizabeth Veinott, for helping me bridge my background in industry to the academic world. Your never ending energy, and your insight and unique perspective have given me great examples of integrating theory and academic research with industry, and I intend to carry those with me as I go back to the real world to further my career.

Thank you to my committee-members, Dr. Leo Ureel and Dr. Laura Brown. I didn't know when I started the program that I would be able to combine my passion for explaining systems to learners with artificial intelligence, but now that I have, I don't want to stop. Thank you for providing me with the foundations to continue my education in AI/ML, and for the guidance and feedback on my dissertation. Your insight and expertise have helped me to grow and have motivated me to continue learning more about AI/ML systems.

To my lab mates, Dr. Kelly Steelman, Dr. Kevin Trewartha, and all of the esteemed professors and staff in the Cognitive and Learning Sciences – any apprehension I had

about starting this program so late in life was wiped away when you welcomed me with open arms. I gladly take the knowledge, experience, and examples you've given me, and I'm grateful for them.

Lastly, thank you to my fabulous husband, awesome sisters, and entire family for your support, and for taking the load off me when I needed it to pursue this dream. People who don't really know me ask me why I'm doing this to myself at this point. But you, who know me well, would've thought something was wrong if I didn't jump at this opportunity, and you've been there to help me throughout the entire journey. I wish mom and dad could be here now that I'm in the homestretch, but I'll settle for the peace and happiness I feel knowing they're in heaven.

Definitions

Artificial Intelligence

Machines and systems simulate human intelligence, they are designed to think and respond to input like humans, and to learn from experience and various forms of guidance. AI systems use algorithms, models and quantitative data analysis to perform tasks that were previously only performed using human cognition (e.g., perception, learning, and problem-solving).

Classifiers

A machine that uses an algorithm to automatically order or categorize data into one or more of a set of "classes".

Intelligent Software Systems

Systems that embody intelligent behavior; could be rudimentary programs, applications, artificial intelligence and machine learning systems.

Machine Learning

A machine that "learns" data and develops expertise on categorizing or garnering that knowledge over time.

Negative Exemplars

Examples showing instances when the classifier did not correctly classify the input.

Positive Exemplars

Examples showing instances when the classifier correctly classified the input.

List of Abbreviations

AI	Artificial Intelligence
CT	Cognitive Tutorials for AI/ML Systems
ERL	Explicit Rule Learning
FSD	Full Self-Driving
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology database

Abstract

Today's intelligent software systems, such as Artificial Intelligence/Machine Learning systems, are sophisticated, complicated, sometimes complex systems. In order to effectively interact with these systems, novice users need to have a certain level of understanding. An awareness of a system's underlying principles, rationale, logic, and goals can enhance the synergistic human-machine interaction. It also benefits the user to know when they can trust the systems' output, and to discern boundary conditions that might change the output. The purpose of this research is to empirically test the viability of a Cognitive Tutorial approach, called Explicit Rule Learning. Several approaches have been used to train humans in intelligent software systems; one of them is exemplar-based training. Although there has been some success, depending on the structure of the system, there are limitations to exemplars, which oftentimes are post hoc and case-based. Explicit Rule Learning is a global and rule-based training method that incorporates exemplars, but goes beyond specific cases. It provides learners with rich, robust mental models and the ability to transfer the learned skills to novel, previously unencountered situations. Learners are given verbalizable, probabilistic if...then statements, supplemented with exemplars. This is followed up with a series of practice problems, to which learners respond and receive immediate feedback on their correctness. The expectation is that this method will result in a refined representation of the system's underlying principles, and a richer and more robust mental model that will enable the learner to simulate future states. Preliminary research helped to evaluate and refine Explicit Rule Learning. The final study in this research applied Explicit Rule Learning to a more real-world system, autonomous driving. The mixed-method within-subject study used a more naturalistic

environment. Participants were given training material using the Explicit Rule Learning method and were subsequently tested on their ability to predict the autonomous vehicle's actions. The results indicate that the participants trained with the Explicit Rule Learning method were more proficient at predicting the autonomous vehicle's actions. These results, together with the results of preceding studies indicate that Explicit Rule Learning is an effective method to accelerate the proficiency of learners of intelligent software systems. Explicit Rule Learning is a low-cost training intervention that can be adapted to many intelligent software systems, including the many types of AI/ML systems in today's world.

1 Introduction and Motivation

This research focuses on training humans in intelligent software systems, which range from rudimentary software programs to more advanced artificial intelligence and machine learning (AI/ML) systems. Intelligent software systems generally embody intelligent behavior and have inner workings that are nested in many layers, each with its own hierarchy and relationships.

The underlying structure, logic, and rationale of these sophisticated systems can be complicated. Although the human learner doesn't need to know the detailed inner workings of these systems, they do need to have some understanding of the system's logic, rationale, and goals. It is also beneficial to have a basic knowledge of when the system succeeds, when it fails, what boundary conditions cause the system to change its output, and what parameter modifications can change the system's output.

One approach that has been used extensively to train humans in these systems is exemplar-based training. Although helpful with specific cases, research has shown that exemplar-based training does not instill accurate global representations of the system, nor does it promote the development of accurate mental models (Smith and Medin, 1981/1999). Learners might need more context and strategic guidance, more information about the relationships of objects and features, along with detailed and summary level cause-and-effect data between the system's inputs and outputs. One way to convey this information to a human learner is with rule-based training. Rule-based training can help

humans understand the comprehensive decision-making process and goals of a system, promoting the ability to apply this knowledge to novel situations in the future. One approach might be to combine the benefits of exemplar-based and rule-based training approaches, which might demonstrate local cases in the context of a global rule, helping learners achieve a more robust understanding of complicated systems. This approach might also support skill transference in a range of contexts.

Using the foundation of Cognitive Tutorials for AI/ML systems (Mueller et al., 2021), preliminary research has been conducted on a novel, non-algorithmic approach to train humans in AI/ML systems. The method, Explicit Rule Learning, provides the learner with probabilistic and verbalizable rules. These are presented to the learner in the form of a rule card which contains textual explanations of the rule, exemplars of when the rule succeeds, the boundary conditions that affect the system's output, and the probabilities associated with the system's successful or unsuccessful application of the rule. The rule card also contains an overview of the rationale and logic behind the system's behavior.

Verbalizable Rules with exemplar-based reinforcement might help learners form better representations of the system's inner workings, resulting in more accurate mental models and better skill transference. These global rules do not aim to explain the detailed, situation-specific workings of these sophisticated systems, and they aren't 100% accurate, but they clearly demonstrate probabilistically occurring outputs that exemplars alone are unable to do.

2 Review of the Literature

2.1 Cognition and Learning

With the learning process, we use our cognitive abilities to make sense of new information, integrate it with our existing knowledge, and apply it appropriately to a situation. The following literature review will evaluate the relationship between cognition and learning, examine theories and research findings from seminal works in the cognition domain, and apply them to inform a rule-based training program.

2.1.1 Learning

Learning has been defined in many ways. Watson (1925) taught us that behavioral changes occur as a result of learning. Thorndike (1908) defined learning in terms of achievements. Klausmeier (1974) defined the steps to conceptual learning as a progression starting at the identity level, where the information is compared to existing knowledge, and ending with the formal level, when the learner can define and be discriminative of a concept. This progression is dependent on the learner's experience (previous knowledge) and the training platform. More recently, Polk (2018) stated that when we learn, we are acquiring knowledge or behavior responses from experience. The consensus among cognitive scientists is that learning is a process in which new information, habits, or abilities are acquired, and which subsequently modify behavior.

We may posit that when we are learning, we are acquiring some type of knowledge and/or experience, and our retention of the matter being learned will help us to modify our subsequent behavior or knowledge of a domain. There is usually a knowledge gain

(*what* we learn, facts, etc.), and a response (changed behavior or enhanced knowledge state). We can positively adjust our cognitive state by taking previously encoded knowledge, incorporate it with new information, and form an updated schema. When we learn, it is desirable to have it persist in our memory, so that we can recall it and hold it up to similar novel situations, iteratively forming new patterns of knowledge and schemas. Learning might be based on information, emotions, or new habits or skills. It generally starts with pieces of declarative knowledge and progresses to procedural knowledge, which is more automatic and discriminative (Anderson, 1982).

There is a vast amount of training methods and presentation modes available. And today's complicated and multi-faceted intelligent software systems require a new type of learning and mindset, with a need to equip learners with multi-dimensional and interconnected hierarchical mental models. Theoretical and practical instruction can be used to build the breadth of foundations, support the learner as they master the depth of a system, and help them to update existing schemas, applying previous knowledge to new circumstances.

An exemplar-based training program provides novices of a system with historical cases; however, it would take many examples, facts, and/or prototypes to educate learners in an intelligent software system. An alternative might be to use rule-based training content, which provides learners with probabilistically occurring patterns. Within the context of rule-based training, exemplars might be used to support the memorization. This

combination might be used to accelerate the proficiency of learners of an intelligent software system.

The focus of this dissertation is to investigate a rule-based training method, complemented with exemplars. This method is being proposed as an efficient and effective technique to train learners of an advanced intelligent software system, enabling learners to achieve more accurate and robust mental models, giving learners a higher level of success in understanding the system's interpretation of the input, with the potential to relate it to the system's output in future novel situations. In this method, the rules provide context, causality, and feature relationships that exemplars alone are unable to provide.

The following literature review explores components of cognition and their role in learning, with a focus on learning intelligent software systems. It also illustrates how rule-based learning complemented with exemplars can accelerate the proficiency of novice users of an intelligent software system compared to exemplar-based training alone. There has been much research on cognition and learning. In that vein, seminal research can provide a theoretical foundation, and more recent applied cognitive research can build on those foundations by providing strategies when training humans in intelligent software systems.

Cognition, as it relates to learning, will be addressed as follows:

- Attention and Perception
- Memory
- Representation, Mental Models, and Categorization
- Judgments, Inference, and Decision-Making
- Generalizing Knowledge to Novel Situations

2.1.2 Attention and Perception

William James (1890) defined *attention* as an active process, whereby our minds take possession of one object, and focus on it with concentration and consciousness of its essence. Concurrently, we withdraw from other objects in order to effectively deal with the attended object.

Perception is the awareness of the object to which we are attending. Using our senses and previous knowledge, along with current goals, we recognize, infer, observe, and discriminate in order to organize and give meaning to the object of our attention (*APA Dictionary of Psychology, n.d.*).

Attention and perception are closely related; our attention is critical to shaping our perception. Attention selectively filters, focuses, and prioritizes information (Kahneman, 1973), and perception interprets and makes sense of the information.

Kahneman (2011) categorized our attention and information processing into two discrete systems: System One is the ultimate goal of skill acquisition: fast, automatic, fluent, and

is performed with ease, without self-awareness or control. According to Kahneman, in general terms, this accounts for 98 percent of our thinking. System Two is slow, deliberate and conscious, effortful, using a self-aware and controlled mental process, rational thinking, and skepticism to seek new or missing information. This accounts for the remaining two percent of our thinking.

The process of skill acquisition, then, is a balance between System One automatic processing, endorsed by System Two insightful feedback, filling in missing relationships, rationale, and declarative facts. For example, in a rule-based training program, the conditions under which the rule applies, and exemplars supporting the rule, might be stored as declarative facts (System Two). Once practiced, a situation with similar attributes might be recognized automatically (System One), and upon endorsement (or disconfirmation) of the rule's application (System Two), the learner applies (or decides not to apply) the learned rule to the new situation.

Similar to Kahneman's System One theory, Dijksterhuis (2004, Dijksterhuis et al., 2006) explored the use of "deliberation without attention" when learning intelligent software systems. Conscious deliberation limits a learner's resources, taxing memory capacity and forcing the learner to consider a subset of the relevant information. This imposes a limitation on learners, who might inappropriately assign weights to features (Kahneman and Tversky, 1982; Levine et al., 1996; Wilson and Schooler, 1991). One might consider that using a conscious attentional effort may result in a higher quality of choice; however, similar to Kahneman's theory that we successfully use the automatic System One

thinking more frequently, in Dijksterhuis's studies, under complex circumstances, participants made better choices with the more automatic form of deliberation without attention.

The ability to selectively attend to some features and objects, while ignoring irrelevant information is crucial for rule-based learning. Rule-based training guides the learner by identifying relevant cues and attributes, directing the learner's limited attention capacity to vital information. Absorbed initially in the form of declarative knowledge, this collection of objects and features to which the learner is attending can be converted to rule-based collections of causal relationships (Hoffman and Klein, 2017; Klein, 2018).

A carefully curated collection of exemplars might instruct the learner by showing them many instances of a pattern. A more robust method, a cause-effect-based rule, considers a collection of objects and features, identifies relationships amongst these, assesses the degree to which each object or feature contributed to the outcome, and assesses the overall richness of the cause-effect without conscious, overwhelming System Two thinking, which would unnecessarily tax the learner's resources.

Previous research found that one way to ensure that learners are attending to the proper objects is to provide the learner with a verbalizable, explicitly stated rule, which will allocate attentional resources to the right target, teaching the learner discriminating factors while guiding them to ignore irrelevant objects. This provides learners with targeted training, developed strategically (Goldstone et al., 2015). This will also assist the

learner by providing a more meaningful collection of objects and features. Exemplars alone might be used to show the learner specific cases, but the onus is on the learner to draw meaningful conclusions and patterns, requiring them to make sense of an arbitrarily grouped set of features, guided by their untrained and possibly anthropomorphized instinct (Mueller, 2020).

Explicit rule-based training, complemented with exemplars, guides attentional resources effectively, and the context of the rule helps the learner perceive the knowledge properly, with the hope of storing it in a retrievable chunk, recalling it in future, similar situations.

The next cognitive component is memory. One way to accelerate the proficiency of learners of a new intelligent software system is make the learned information memorable, which increases the likelihood of it being encoded as a memory (Zhang, 2019).

2.1.3 Memory

Human memory can be divided into three processes: encoding, storage, and retrieval (Baddeley and Logie, 1999). A fourth process might be application, which is knowing when to apply an encoded memory. Effective encoding can lead to more meaningful and efficient recall (Roediger and Goff, 1999).

Exemplars can benefit a learner by providing vivid snapshots of cases that can be ingrained into memory through repetition. The quantity of exemplars, and the necessary repetition might expose the learner to innate human memory limitations (Miller, 1956), requiring the learner to self-identify relevant objects and features, and create meaningful

categories and patterns autodidactically. This might be challenging with intelligent software systems that have many interconnected components and complicated relationships. Furthermore, an effective combination of short-term and long-term memory is important when learning (Nickerson and Adams, 1979). An exemplar-based training program might overload working memory, while the learner tries to analyze the exemplars and identify patterns and differences, resulting in unclear retrieval from long-term memory, from which the learner is using resources to match new exemplars to already known categories and patterns.

However, rule-based training can ease this resource overload, by providing the learner with a clear framework for organizing and categorizing objects and features. An effective rule can demonstrate the underlying organization of the system, provide context (when/where to apply the rule), making the training more meaningful. It can also provide the learner with memorable retrieval cues, by applying rationale and principles to a collection of objects and features. This can also help later, when the learner needs to generalize this knowledge into new, unseen situations (Rasmussen, 1983). Rule-based training can be more functional, which Nickerson and Adams (1979) found to be more memorable.

The causal nature of rules can also provide a deeper understanding of concepts (Craik and Lockhart, 1972), providing a more elaborate but concise and structured short-cut, as opposed to many declarative exemplar-based instances. Rule-based training provides learners with nondeclarative knowledge that can be used to explicitly learn relationships

between features, objects, and events; learners can extract common elements from a series of separate combinations of objects, features and events, rather than individually presented exemplars (Squire, 2004).

When learning an intelligent software system, much of the information processed by the learner is complicated and ambiguous. Using rule-based training provides a distinctiveness of input/output combinations, something shown to benefit encoding (Roediger and Goff, 1999), which can be stored and later retrieved and applied to new, similar situations.

Once information has been stored in memory, it needs to be functional, so that it can be retrieved, instantiated, and applied to a situation, based on the learner's broader mental model of a system.

2.1.4 Representation, Mental Model, and Categorization

Learners of a new system take previously encoded knowledge, process new information with it, and form a mental representation in their mind (Bruner, 1964). The representation of knowledge used by a learner can provide building blocks (or stumbling blocks) for their mental model and can have severe implications on their ability to solve problems (Amarel, 1968). These representations accurately depict encoded information and are also essential for learning new concepts and skills, as they are used to organize and interpret information in a new system.

A mental model is a framework, an organized system of concepts that an individual forms as they encounter a system (Mueller et al., 2021). It is the basis of the individual's understanding of a system, an organization of the features of the system and their interconnectedness. It contains relevant concepts, relationships, and causal factors (Johnson-Laird, 1989).

Having a good mental model helps learners retain information longer (Kieras and Bovair, 1984), and can be used diagnostically to identify potential gaps in knowledge (Bravo-Lillo et al., 2010). A good mental model of a system's features and structure gives the learner a strong foundation from which they can retrieve relevant information to mentally simulate and solve/predict future, unseen problems. Research has shown that an expert's mental model can differ from a novice's (Chi et al., 1981), with the former being more effective. An expert's mental model is not just a static representation of information, but a structure which can be manipulated; features and objects can be mentally modified to simulate various outcomes. A learner might simulate an analogous outcome (Gentner, 1983), or a different outcome altogether, by manipulating features and objects beyond boundaries and limits. In Klein's recognition-primed decision making (RPD) model, he suggests that experts' mental models are an essential asset that helps them recognize patterns in novel situations, enabling them to make rapid and effective decisions, even in exigent circumstances.

Mental models can be robust (Chi et al., 1991), and can be formed with well-developed training material (Hitron et al., 2019; Mueller and Klein, 2011). Rule-based training can

help a learner with their representation and mental model. Rules can provide a structure for organizing information, helping the learner to make sense of large amounts of data, and compartmentalize cause-effect sequences into more manageable and easier to remember chunks (Ashby & Gott, 1988). Rules also show the learner the combination of objects and features that make a difference in a system's output, helping the learner to construct a mental model that includes relationships and, when appropriate, causality. Lastly, rules can help a learner see similarities between a known input-output combination and a new situation, drawing parallels, enabling transference.

2.1.4.1 Categorization

One way learners represent information is in the form of categorization, where individual concepts are part of a larger, more comprehensive system, containing similarities, patterns, and boundaries (Vosniadou and Ortony, 1989). Categories allow learners to organize and understand new information by providing maximum information with the least cognitive effort (Rosch, 1978).

Categories might be simple or complex (Medin and Smith, 1984). For example, a simple category might be "boy", and a complex category might be "rich boy". Furthermore, there are different levels of abstraction. Murphy and Brownell (1985) found that in most cases, participants responded more quickly when presented with a "basic" level of abstraction (e.g., car) versus the "superordinate" level (vehicle) or "subordinate" level (sedan).

When learning an intelligent software system, an exemplar-based training might show individual examples of the basic level of abstraction, which might be similarity-based averaged-shaped objects (Rosch et al., 1976). These might be prototypes (Homa et al., 1973; Posner and Keele, 1968) or may have a collection of representative features from which the learner infers category membership (Fried and Holyoak, 1984; Murphy and Medin, 1985). As previously noted, it is often difficult to determine a learner's criteria for similarity or dissimilarity ultimately used to assign one category over another (Smith and Sloman, 1994), especially since perceptions, previous experiences and knowledge may differ (Tanaka and Taylor, 1991). These factors are critical to the success (or failure) of a novice's interaction with an intelligent software system, especially if the learner is using Kahneman's System One (fast) thinking, or Klein's recognition primed decision making (Klein, 1993), which are intuitively based upon their mental model. An accurate mental model will serve the novice well, an inaccurate mental model can lead to errors.

A category might be "ad hoc". This is not a classical categorization in the sense that there is a known, learned collection of items in a category (e.g., a "fruit" category might contain "apple", "orange", "fig", etc.). Ad hoc categories are formed to achieve real-world goals, and are dependent on the learner's understanding and representation of real-world situations (Barsalou, 1983; Edinger and Goldstone, 2022). As an example, an ad hoc category might be "things you take to the beach" (S.T. Mueller, Memory and Learning class, October, 2019).

A learner of an intelligent software system being trained with an exemplar-based training program might form an ad hoc category for a collection of examples. These might be created summarily, using naively formed representations and mental models as a basis to make sense of the examples, which would subsequently be generalized to novel situations; this may lead to erroneous results.

Categorization plays an important role with regards to a human learner's representation and mental model. However, it is also a fundamental concept when describing the inner workings of many intelligent software systems. It is an integral part of the rationale, logic, functionality, and output of these complicated systems. For example, an image classifier might categorize pictures into categories of "dog" or "cat". A loan system might categorize application into categories of "accept" or "reject". Furthermore, an advanced intelligent software system can use categories to explain its reasoning and rationale, expounding on the features and variables that were integral to its output. Using categorization as a framework makes these advanced systems more transparent and understandable to humans. Many intelligent software systems determine category membership by using rules. Concurrently, many humans are trained in intelligent software systems with category-derived, rule-based training.

A rule-based training program takes a collection of exemplars, applies explicit rules with known logic, rationale, and probabilistically occurring outcomes, leading the learner to a better understanding of the system. An effective rule-based training program also relates an intelligent software system's category logic and functionality to the learner.

When learning an intelligent software system, certainly some categorizations might be more benign, with minimal adverse effects and insignificant consequences (e.g., an image classifier misclassifies a cat as a dog). Others might have more significant consequences. For example, a new driver of an autonomous vehicle needs to be able to predict the possibility of a dangerous situation and be ready to make a decision to let the vehicle proceed autonomously or disengage the autonomous system and take over control of the vehicle. One event might be categorized as "the lane lines are clearly visible, so the autonomous vehicle system *will drive as expected*, on the right side of the road". Another event might be "there are not clearly painted lane lines, so the autonomous vehicle *will drive unexpectedly in the center of the road*, straddling the oncoming lane and the right lane. In this case, the categories used by the intelligent software system are "lane lines = drive (properly) in the right lane", "no lane lines = driver (erroneously) in the center of the road, straddling the oncoming lane and the right lane". The categories for the driver are "let the vehicle continue operating autonomously, there is no danger" or "disengage the system and take control of the vehicle, potential for danger."

Rule-based learning is explanation-based, goal-based, top-down, and is guided by expectations and experience. Chomsky (1980) proposed that learners have an innate ability to apply learned rules to new situations. Incorporating exemplars into rule-based training programs might provide learners with a better opportunity to form more accurate mental models. Kahneman (2011) proposed that initially, System Two thinking (deliberate and rule-based) develops mental models that are more accurate and robust

than System One (intuitive and automatic) thinking. When the learner achieves a higher level of understanding, they progress to the more intuitive and automatic level, at which point the exemplars can be recalled in the context of the rule.

Exemplar-based learning is similarity-based, bottom-up, guided by the perception, representation, and mental model of the learner, whether accurate or not (Medin and Heit, 1999). Categorizations might result in desirable outcomes, or deleterious events. Exemplars within the confines of a verbalizable, probabilistically occurring rule might ensure a better outcome.

2.1.5 Judgments, Inference, and Decision-Making

Sound judgment in learning leads to a more effective evaluation of the learned information. It utilizes critical thinking and helps the learner make informed decisions about the accuracy and relevance of the information, as well as the information's relationship to existing knowledge. Inference is the ability to draw conclusions or make categorizations based on the information provided. This includes identifying patterns, boundary conditions, exclusions of category membership, and relationships, all which help the learner solve problems and promote the development of an accurate mental model.

In the previous section, learning by using categorization was reviewed. We will further explore categorization by identifying the effects of judgment and inference on a learner's ability to proficiently categorize an intelligent software system's objects and features.

When a learner encounters new information, it is evaluated not on its own, but in the representativeness and context of existing knowledge, as well as past outcomes, both in similarity and category representation (Kahneman and Tversky, 1973). This might be effortful and conscious, or it might use "adaptive unconsciousness", which are mental processes that work rapidly and automatically using relatively little information (Gladwell, 2006). When presented with new information, the learner makes judgments on the inclusion or exclusion of membership to an existing category. If the degree of similarity is meets a certain threshold and is high, or if the basic level of abstraction matches, it is considered a "fit" into the existing category (Osherson et al., 1990).

Sometimes the new information is ambiguous. For example, what if the new information matches the features of one category, but the object is prototypically similar to another category? The learner will make inferences, and many times conclude that the prototypicality of the category has more weight than the features (Gelman and Markman, 1986). For example, a whale looks like a fish, lives in water, and swims like fish do; however, it is not a fish, it is a mammal. A learner considering these features may erroneously conclude that a whale fits the prototype of a fish, rather than considering the properties that make it a mammal. In any event, once this new information has been categorized, it will follow the schema for that category, whether that be the functionality associated with that category, the decisions associated with that category, etc.

Humans are bias- and error-prone in judgments, inferences, and decision-making (Tversky and Kahneman, 1974). The quality of these judgments, inferences, and

decisions can differ with varying levels of expertise (Ericsson and Kintsch, 1995). When inferring information from one or more objects and features, humans have an erroneous tendency to recall the inferred information at a later point in time, as if it was actually presented to them declaratively (Polk, 2018).

This bears heavily on how different training methods help learner proficiency. A learner needs to be able to make inferences based on stored exemplars as related to rules. A novice learner of an intelligent software system may not have the foundational knowledge to infer proper category allocation with exemplars alone (Smith, 1995); having exemplars in the context of a rule might help to alleviate this error-prone tendency. Rule-based training complemented with exemplars bypasses a novice learner's naive induction, inferences, and bias-laden decisions. The best rules would include exemplars in the context of an if…then statement and offer clear and relevant input/output combinations that are probabilistically likely to occur (Grice, 1975). Factual and counterfactual exemplars strengthen and fortify the learning experience, enabling the learner to achieve a higher level of foundational knowledge, making the learner more discriminate in novel situations.

2.1.6 Generalizing Knowledge to Novel Situations

Generalization (i.e., skill transfer) is a cognitive process in which a learner applies their existing knowledge, concepts, or previously learned skills and past experiences to a new, similar situation (Gick and Holyoak, 1987). The learner analyzes a novel situation, compares it to previous situations, identifies key features and similarities between the novel problem and a previous problem, and applies a previously proven solution to the

novel situation (Holyoak and Morrison, 2005). The degree of variability and processing used in the previously solved problems may make the transfer of knowledge and skills more simpler or sometimes, more difficult (Gentner, 1989).

Analogies also facilitate generalization (Gentner et al., 2001). The very nature of an analogous comparison likens objects and features of one situation to another. In a study, Gick and Holyoak (1983) showed that participants who compared two analogous stories were more likely to generalize their knowledge to future situations. In an analysis of the trade-off between analogies and rules, Forbus et al., (2020) argue that both can be used for skill transference. The flexibility of analogies might be more desirable than the rigid and explicit set of conditions necessary for a rule's application. However, domain-specific rules might be more precise and consistent than a misapplied analogy.

One factor in generalizing previous problems to novel problems is categorization, which, as previously stated, establishes patterns, boundaries, similarities, and exceptions. A structured collection of features and objects, whether learned from exemplars or rules, can help the learner form an effective mental representation, properly structured in such a way that this framework will be generalized in the future to a similarly framed novel problem (Erickson and Kruschke, 1998).

When learning an intelligent software system, learners using an exemplar-based training program might be better equipped to generalize one category to a novel situation, especially if the exemplar is associated with specific context or situations. The contextual

information surrounding the exemplar provides a type of mental anchor. In this way, exemplars are specific and concrete representations of a case-based situation (van Gog et al., 2019).

However, a rule-based training program might equip the learner to better generalize different nuances, or instances and variabilities of a category. In this respect, the learner uses rules to identify patterns, system logic and rationale, and principles, and can help identify the boundaries of categories (Smith and Sloman, 1994). Rules complemented with the support of counterfactual or contrastive exemplars and scenarios strengthen the learner's ability to generalize existing schemas to novel situations. This higher-order thinking encourages abstract reasoning and critical thinking, providing a deeper level of processing.

This combination of exemplar-based training and rule-based training might be the best combination to give learners of an intelligent software system the foundation needed to generalize previously encoded problem/solution combinations to novel problems. Verbalizable rules provide learners with a system's logic and rationale, and guides their attention to the relevant objects and features. This might be complemented with exemplars (factual and counterfactual). Explicit Rule Learning, the focus of this dissertation, investigates the possibility of using this combination to help learners become more proficient in intelligent software systems, and aims to identify whether or not it facilitates skill transference.

This section explored cognition as it relates to learning. Specifically, for this dissertation, the focus is on training humans in intelligent software systems. The goal is to accelerate proficiency and equip the learner with tools to be more discriminate of when the system performs as expected and when it fails, and the conditions that cause it to succeed or fail. One form of intelligent software system, which will be the platform used to research the Explicit Rule Learning training method, is Artificial Intelligence/Machine Learning. The next section will discuss the unique nature of these systems, and methods that have been, and are currently being used to train human-learners of these systems.

2.2 Learning AI/ML Systems

There are some differences between classical school or experiential learning and the learning of Artificial Intelligence or Machine Learning (AI/ML) systems. AI/ML systems are data-centered, dynamic, and interdisciplinary. This calls for different approaches in training methods. The next sections will define AI/ML systems and review traditional methods that have been used to train humans in these systems.

2.2.1 AI/ML Systems Defined

Artificial Intelligence (AI) machines and systems simulate human intelligence; they are designed to think and respond to input like humans, learning from experience and various forms of guidance. AI systems use algorithms, models, and quantitative data analysis to perform tasks that were previously only performed using human cognition (e.g., perception, learning, and problem-solving). AI systems are omnipresent in today's society. For example, you may be sharing the road with a self-driving vehicle, or using a digital assistant such as Apple's Siri, or Amazon's Alexa. Perhaps you have an iRobot to

vacuum your space, or you may have seen Diligent Robotics' Moxi, which delivers medication to patients and samples to medical laboratories.

A subset of AI is Machine Learning (ML), which "learns" from the data it is given and develops expertise on categorizing, cumulatively collecting that knowledge over time. There are three types of machine learning: supervised (where a machine is given input of labeled data and examples of expected results), unsupervised (no help is given, "learns" on its own), and reinforcement (the machine receives no training, but is given positive or negative reinforcement from humans to improve and refine its output).

An example of a reinforcement system is an autonomous vehicle that responds to input and is "corrected" by the human driver supervising the system when the human disengages or grabs control of the vehicle preemptively to avoid an adverse event. When the human takes over control in these situations, the autonomous vehicle's AI system is learning that this set of features and circumstances resulted in a negative (undesirable) output and will consider this information when presented with a similar situation in the future, hopefully leading to more appropriate actions aligned with the human's expectations.

AI/ML systems are used in many domains, from finance to medicine, from the transportation domain to academia. For example, medical diagnostic imaging results might be input into an AI/ML system that's been trained on specific diagnoses, symptoms, or some other medical facet. After the AI/ML system has been sufficiently

trained, it receives data about a specific patient, provides diagnostic and prognostic information, and makes predictions about the trajectory of the patient. Repeatedly processing this data iteratively with data from many patients and providing the system with a proper feedback loop will improve the system's output, increasing the reliability of the diagnostic, prognostic, and predictive accuracy over that of a human (Gillies et al., 2016).

Research has shown some tasks are better left to the human, and others to the AI/ML systems. For example, finding patterns and trends in large, structured data sets in a nanosecond is a great asset of AI/ML systems, much more accurate and efficient than humans, given the same amount of information and time. However, humans are better at considering context, using intuition, and finding patterns in unstructured data.

The research in this dissertation uses two AI/ML systems: an ML image classifier and a more advanced AI-based autonomous driving system.

Next, explainable AI will be discussed. This is one form of helping learners to understand intelligent software systems. Many of the methods used in explainable AI are directly relevant to training new users of AI/ML systems.

2.2.2 Explainable Artificial Intelligence (XAI)

The availability and advances of these AI/ML systems have improved the quality of computing and we are able to receive data on a much higher level than in the past. However, with this phenomenal advancement comes a question: How can we take these

intelligent software systems and help humans to interpret their output in such a way that the human understands the reasoning and decision-making of the system, and can trust (or decide not to trust) the correctness and validity of the output?

A rudimentary, non-AI/ML system, such as Microsoft Excel or accounting software is developed to solve specific problems based on predefined rules and logic. It is pre-programmed with thousands of lines of code, with specific logic, decisions, and flow, following explicit rules. A system that is given the same input, in the same sequence will consistently render the same output. Humans can be trained in these systems and have the potential to consistently predict and rationalize the output.

Training humans in AI/ML systems, however, is different from training them in software and applications where the underlying algorithms have been predetermined and programmed. AI/ML systems are more complex, abstract, and dynamic. Researchers are actively working on interpretability techniques to make AI systems more transparent. Explainable Artificial Intelligence (XAI) is a domain created for this purpose, to explain reasoning and decisions made by an AI/ML system to the human, who can then understand, believe, and optimally trust in the output from the system. In addition to knowing the logic, rationale, and goals of the system, the human also needs to be aware of a system's strengths and weaknesses, as well as the boundary conditions, and the variations to features and objects that change the output of the system.

Explanations of AL/ML systems might be algorithmic or non-algorithmic. An algorithmic explanation is created by the AI/ML system itself, whereby the system performs a post hoc analysis and identifies specific factors or variables that contributed to its output. It communicates the rationale it used to form its output to the human using the system.

A non-algorithmic approach might use natural language communication with the human to explain the variables considered, and the rationale used. For example, after an output, the humans might ask the system questions about the factors and weights that led to that output. One non-algorithmic method is collaborative filtering (e.g., a recommendation engine), which makes recommendations to users based on their previous actions or preferences. Another method might be done manually, where an expert in the domain might interpret the variables and form an explanation for the human user of the AI/ML system.

Whether algorithmic or non-algorithmic, there are many challenges in creating effective explanations for human learners of a system. For example, the level of depth might sacrifice accurate explanations so that the explanation could be more transparent and understandable to a novice learner. Ultimately, explanations should provide more than input-output rationale and logic; they should provide information on goals of the system (hopefully aligned with the goals of the human), and the context in which the explanation applies.

2.3 Human Classification and Categorization

As previously stated, humans use categorization to make sense of the vast amount of information they encounter daily: to organize information, to perceive and recognize objects or patterns, to facilitate meaningful communication, problem-solve, make decisions, generalize information, and make inferences, and of course, when learning new concepts. Using categories when learning new material helps with organization and structure, chunking information, applying existing knowledge to discriminate, differentiate, or identify similar relevant patterns and principles, helping to make predictions and inferences with new information, and organizing and structuring new information into existing schemas. Categories help us to be more efficient and can make information coherent and meaningful.

One way to teach categories is with exemplar-based training. This will be discussed next.

2.3.1 Exemplar-Based Training

Many forms of explanations and training have been researched in order to discover the most effective way to train novice users of AI/ML systems. These include lectures, presentations, demonstrations, hands-on exploration, white papers, guided group or individual learning, case studies, expert-guided learning sessions, and the list goes on and on.

One method that's been used extensively in classical training is Exemplar Training (Smith & Medin, 2002). Smith and Medin (1981/1999) defined exemplars as "members

of a category that are experienced and stored in memory", and these concrete and specific examples are used to represent a larger category. Based on the similarity of the exemplars to new instances, we can make judgments and predictions in order to make decisions or solve a problem.

In this method, learners are presented with typical examples of similar cases (including both prototypical and marginal cases; Mueller et al., 2019) or situations where the outcome or categorization were the same. Cumulatively, these exemplars help to define the categorical logic and reasoning to the learner. Optimally, these examples also demonstrate the boundary conditions that may or may not change the outcome.

Exemplar training might be in the form of prototypes (Homa et al., 1981), where the central tendencies, average, or best exemplars are provided to the learner. Concepts, categories, or typical features might be shown or textually described to the learner (Nosofsky et al., 2018; Smith, 2008). Another way to use exemplars is to show positive (factual, correct outcome) or negative (counterfactual, incorrect outcome) examples of a system’s output (Ohlsson, 1996).

Lastly, exemplar-based training might show the learner worked examples (Kalyuga et al., 2001), where learners are given solved problems in which the step-by-step solution is demonstrated (van Gog et al., 2019). This also benefits the learner by providing an expertly guided structure to be used when solving similar problems.

Exemplar-based training helps learners of AI/ML systems develop better recognition skills, allowing them to discern patterns and group like-items into categories (Mozannar et al., 2022). It also provides for more intuitive predicting, which Kahneman and Tversky (1973) believed were based on the representativeness of past outcomes. Examples can be easily retrieved from memory (Nosofsky and Little, 2010); however, the context and application of the examples could have good or bad results, based on the accuracy of the learner's mental model.

Although commonly used to teach learners of AI/ML systems (Hase and Bansal, 2020; Kim et al., 2016; Krause et al., 2018; Nushi et al., 2019; Poursabzi-Sangdeh et al., 2021; Wang et al., 2016), exemplar-based training has had mixed results. Examples alone may not be enough to explain the principles, algorithms, and goals of the system, and their scope is limited to the quantity and scope of the examples presented to the learner. Learners with different levels of experience and expertise in a domain or with AI/ML systems may differ in their interpretation, categorization, groupings, or generalizations of the system's functionality, which could help or contaminate their learning.

2.3.1.1 Problems with Exemplar-Based Training

Although exemplar-based training may make learners more proficient in some domains, when considering advanced, unstructured intelligent systems, it is lacking. In these situations, the learner needs to autodidactically discover which features and objects are important, the weight of each of these, and they must interpret the hierarchy, relationships, and any causality factors on their own.

Additionally, exemplars alone, even when they include counterfactual and contrastive examples, need to be presented with an appropriate number of examples, strategically selected, so that they demonstrate to the learner the categories which are to be learned (implicitly and explicitly). There should be enough examples to demonstrate a distinct pattern; however, there cannot be too many examples so that the learner is inundated with too many patterns and categories to remember.

2.3.2 Rule-Based Training

Another way to present categories and information to learners is via rule-based training (Lakkaraju et al., 2016; van der Waa et al., 2021). This method uses executive, higher-order cognition to support the evaluation of a combination of objects, their features, and any relationships and causality. A rule specifies the necessary and sufficient conditions which determine its membership in a category (Smith et al., 1998). This is a more global and explicit form of training that is organized into meaningful chunks, offering "if...then" statements, indicating combinations of, and relationships between conditions/variables, context, and the likely outcome. With this type of instructional strategy, the learner evaluates certain specific features (input), applies them to a previously learned combination of input/output and context, and decides if the new item belongs in a category (i.e., is a similar situation).

Rules should be simple (Grice, 1975; Jung et al., 2017), and based on explicit reasoning that treats each feature/object dimension separately but should also explain the relationship between them (Jung et al., 2017) as well as the aggregate meaning of the collection.

2.3.2.1 Rule-Based Training vs Exemplar-Based Training

Rule-based training is different from exemplar-based training (Nosofsky et al., 2018). For example, an exemplar-based training might give the learner a declaratively presented flock of "birds" (e.g., eagle, sparrow, hummingbird, etc.). A learner might memorize this group, encode it, and infer categorical membership with a new unseen object. However, a rule-based training program will provide if…then conditions under which an object is a bird (e.g., if wings=true && feathers=true && hasBeak-Bill=true && mammal=false, etc., then "bird"). Rules may not always be 100% predictive of categories, but rather, they're a useful way to shortcut the sensemaking (Klein et al., 2007) process about a system.

Wolfe (1994) differentiates bottom-up and top-down visual search processes. When evaluating features, evaluating exemplars is a bottom-up process, where the learner is comparing features of nearby objects, looking for distinctions and patterns. However, rules are a top-down, user-driven evaluation of features, driven by goals and strategies. In this respect, rules can help identify collections of features that matter, as well as the expected outcome.

Exemplars are an excellent tool that can be used as support for rules. Many exemplars could be distilled into one verbalizable statement that helps the learner better able to predict system performance and categorize information (Paul and Ashby, 2013). When the learner is recalling the rule, exemplars are recalled, either as valid applications of the rule, or as exceptions (Rutledge-Taylor et al., 2012). In this respect, concrete, vivid

exemplars can be the anchors relied upon to trigger a previously learned rule and measure the rule's relevancy to a new situation. Ultimately, although the exemplars support the rule, the rule's conditions and boundaries are the key that enable the learner to generalize learned material to new situations.

2.3.2.2 Benefits of Rule-Based Learning Complemented with Exemplars

A benefit of rules, and their positive contribution to learners' perception, memory, representation, mental model, categorization, inferences, judgments, and skill transference is the rationale and causal reasoning encapsulated in them. Not all rule-based training provides causal reasoning; some rules might identify correlations and associations between features and objects without identifying the causal relationships between them. However, there are benefits to rules that demonstrate causality. For example, in the context of autonomous driving, a causal predictor might be that if the vehicle's sensors and cameras detect an object in the vehicle's trajectory, it will brake to avoid impact.

Causal reasoning is central to the learner's sensemaking of the system. It can improve their mental model, guide decisions, and help the learner adapt to dynamic situations, knowing when to pause and re-plan the direction they're going in, and in the coordination between the human and the intelligent software system. It can also help their ability to anticipate the system's likely outcome with/without intervention by the human.

Probabilistic rules inform the learner about the likelihood of events or outcomes. Generally, these rules have an antecedent, which is the observed features and objects (input), a consequence, which is the likely outcome (output), a probability, which is the likelihood of the consequence given the antecedent, and parameters, which is the source of the data (i.e., the sample) from which the antecedent and outcome were used. As previously stated, it may not be possible to form rules that are 100% predictive, but probabilities help learners make better categorizations and decisions than they would in situations of uncertainty. The goal with using probabilistic rules is to help learners understand the underlying principles and rationale used by the intelligent software system, and to provide the learner with a framework and ability to forecast the most likely outcome.

One way to make rules effective for novice learners of an intelligent software system is by making them verbalizable (i.e., explicit, using memorable statements). This should help the learner to encode the information into memory more effectively, making it easier to retrieve and apply to new situations in the future (Maddox and Ashby, 2004). Verbalizable rules demonstrate to the learner what objects and features to focus on (Mueller and Weidemann, 2008), and can be supported with factual and counterfactual exemplars. Exemplars alone (i.e., not in the context of a verbalizable rule) leave the learner to come up with their own categories, which could be clouded with noise the learner may naively consider when making categorizations.

One method that has been developed to train learners is Cognitive Tutorials for AI/ML Systems. This approach will be described next.

2.4 Cognitive Tutorials for AI/ML Systems

Cognitive Tutorials for AI/ML Systems (CT) apply cognitive principles (i.e., human strengths and weaknesses with regards to attention, memory, decision-making, etc.) in their approach for training people in intelligent software systems such as AI/ML systems.

Explanations of AI/ML systems might be local or global (Mueller et al., 2019). Local explanations are post hoc and explain why a system had an output in a specific case, and which features and objects contributed to the output. The assumption is that these cases occur with some regularity, and that by learning these cases, the user of these systems will be able to recognize them in the future and apply them accordingly. Global explanations are meant to convey the overall understanding of how a system works, identifying important features that contribute to the model's output. Although local explanations help learners understand specific details of a system's output (Mueller et al., 2021), global explanations can contribute to a more robust mental model, resulting in a better understanding of the overall rationale and logic of the system.

Cognitive Tutorials are global explanations that formalize, document, and train humans in cognitively challenging systems. Cognitive Tutorials are helpful for learners of intelligent software systems. They help humans acquire a more functional and accurate representation and mental model of the system; they are a means to accelerate the

development of a learner's understanding of a system as well as the learner's proficiency in predicting and understanding the system's output.

The Cognitive Tutorial approach considers an array of system attributes, such as: the system's data requirements, representation modes and modeling mechanisms, underlying computations and algorithms, and the output form of the system. These are used to inform tutorials, which might be presented in the form of: walkthroughs, forced-choice scenarios, troubleshooting/inducing errors, novel problem presentation coupled with an expert's solution, rule training/untraining, counterfactuals contrasts, semi factual-counterfactual sequences, mental model matrices, cheat sheets, or the shadowbox approach (Klein et al., 2013).

2.5 Explicit Rule Learning for AI/ML

One type of Cognitive Tutorial that is the basis of this research is called Explicit Rule Learning for AI/ML. This method begins by presenting the learner with a probabilistic and verbalizable rule, factual and counterfactual exemplars (which illustrate the boundary conditions), the probability of occurrence, and a summary of the rule's effectiveness. This is presented to the learner on a "Rule Card". Next, the learner is given real-world practice situations where they apply the rule. This is called "Practice with Feedback." After each practice response, immediate feedback is given to them, either confirming their proper application of the rule if they answered correctly, or, if the learner's answer was incorrect, they receive a description of why and how the correctly applied rule should have been used.

The Practice with Feedback component increases the learner's level of engagement. Several frameworks have been developed regarding the benefits of increasing a learner's level of engagement. Craik and Lockhart's (1972) Levels of Processing framework states that deeper processing (e.g., elaborating on the material being learned) will result in higher retention.

Similar to this framework is Bjork and Bjork's (2011) usage of desirable difficulties, whereby introducing more difficult material leads to stronger encoding and retrieval. By engaging in active and varied practice, being tested on the material, generating answers, and practicing retrieval of the learned material, learners engage in enhanced encoding and deeper processing of the information, which leads to a higher level of proficiency on the material. Introducing a higher level of difficulty and variable context to the practice and testing stages encourages the learner to process the information deeper, and to better adapt the learned material to new situations.

The Rule Card and Practice with Feedback components follow a standardized format, which help ease the cognitive load. Van Merrienboer and Sweller (2005) found that complex learning was optimized when cognitive load theory was considered, and the cognitive capacity of learners was managed. Reducing extraneous information and stimuli, and focusing on the proper features and objects, and application of the rule in a standardized and focused manner with Explicit Rule Learning enhances learning.

Explicit Rule Learning for AI/ML is an active learning process. Learners transfer rule knowledge to new situations, continually improving schemas, giving scenarios more meaning, which help the learner to better understand the system's rationale and justification, and this feeds back to help the learner expand their mental model. This exceeds the proficiency potential when compared to typical help manuals. Typical help manuals are used as "just-in-time" tools, giving the learner text or images to read based on keyword search, with the hope that this will help them solve their current problem. In contrast, rule-based training offers deeper levels of processing, and builds skills that are global, and more transferable to novel situations. Intelligent software systems are complicated, and rules provide visual examples, context, causality, logic, and rationale, which single-dimensional help manuals are unable to provide.

2.5.1 Explicit Rule Learning Steps

The following are the steps used to create an Explicit Rule Learning training program:

1. Identify relevant rules: The goal is to identify rules which, as a collection, will aid the learner in developing a better mental model of the entire system, the system's rationale and logic, illustrate to the learner important features and objects, and help the learner to understand the overall framework, tendencies, goals and strategies of the system.
2. Provide the rules to the learner: These are presented on a "Rule Card" (see Section 2.5.3.1 for more detail).
3. Practice with Feedback: The learner progresses through real-world situations, with the goal of applying the proper rule appropriately to each situation.

4. Feedback: The learner is given immediate feedback on the correctness of their response, stating whether or not they answered correctly, with a description of which rule applied, and the rationale behind the specific rule's application to the current situation (see Section 2.5.3.2 for more details).

2.5.2 Rule Properties

The rules that are selected for the training have the following properties:

- They should be verbalizable, global, and clearly stated
- They should contain a description of the most likely output resulting when the rule is applied
- The possible reasoning or logic for the system's output should be identified
- They should include feature salience descriptions and weights
- The context in which the rule should be applied should be identified, along with applicable assumptions or constraints
- Boundary conditions of the rule should be identified, along with examples of possible modifications to the input that may result in a different output
- Should be generalizable to future situations

2.5.3 Explicit Rule Learning Components

ERL has two components: a Rule Card, and Practice with Feedback. The following is a more detailed description of these.

2.5.3.1 Rule Card

This one-page description of an explicitly stated, verbalizable rule can be presented in a printed or digital format. It contains a textual explanation of the rule, visual examples

(exemplars) of cases when the rule applies (factual), and the conditions under which it may not apply (counterfactual), a probabilistic value of the rule's sensitivity, and a textual summary of the rule's effectiveness.

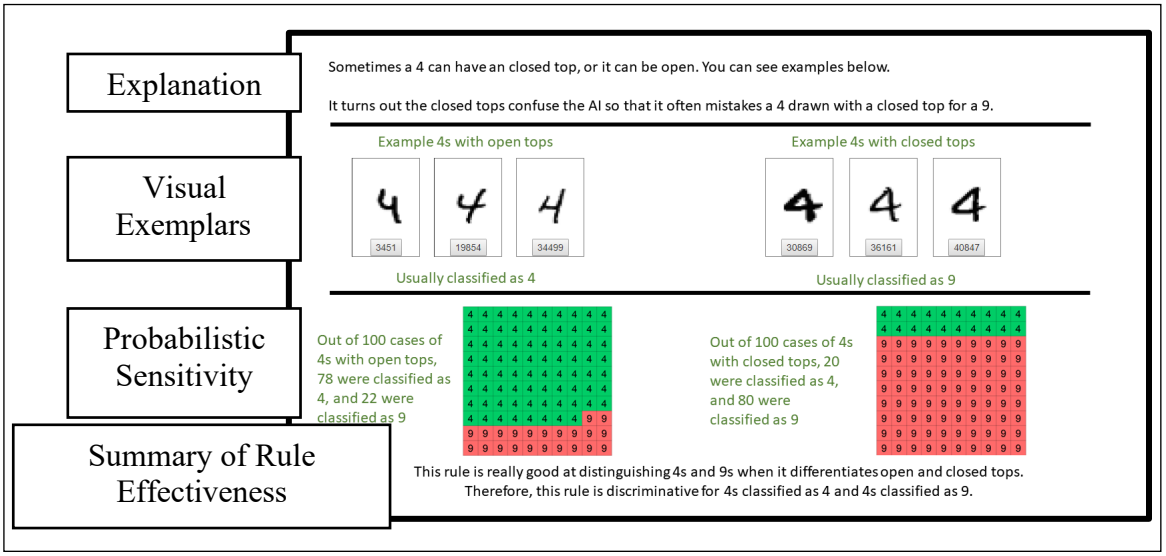
This at-a-glance, portable or digital card is a concise representation, and uses system nomenclature (defining new terms as necessary), key concepts, and processes the system uses (Figure 2.1). It contains factual and counterfactual exemplars that will later be recalled in conjunction with the rule. A Rule Card is also handy as a tool to refresh a learner's memory as they refer to it on an as-needed basis in the future. It provides the learner with structure, a clear and concise representation of the problem space, and the system's framework.

The benefits of the rule card to the learner are as follows:

- Quick access to important information condensed into an if…then statement
- Improved retention (verbalizable, with visual exemplars, textual explanations, and probabilistic information)
- Reduce cognitive load by providing an easy-to-use reference of the important features, objects, information (Norman, 1988)
- Consistency: rule cards can help prevents errors and increase accuracy by providing a formal, standardized method of delivering rule-based training

Figure 2.1

Sample Rule Card for an ML Image Classifier

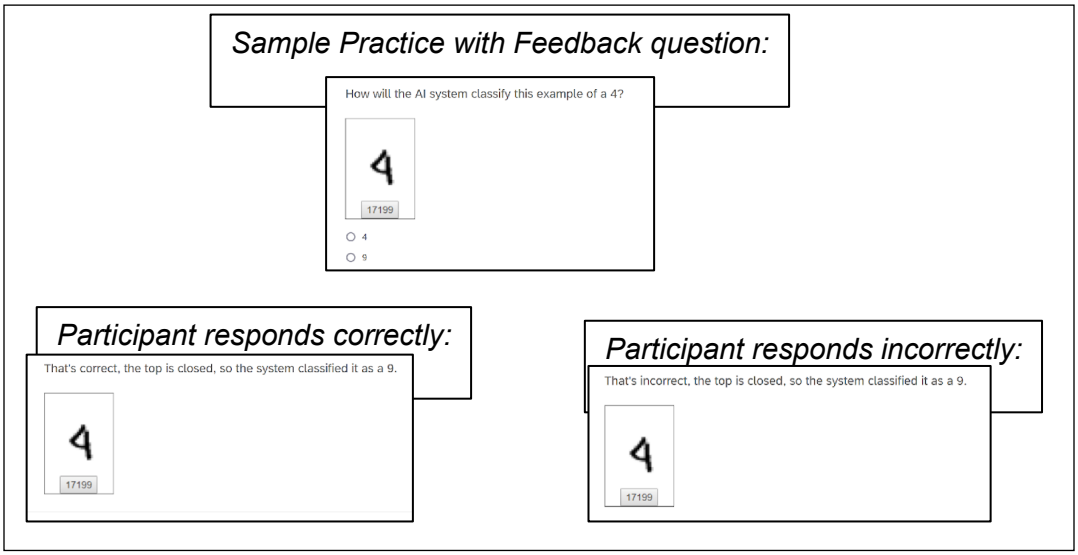


2.5.3.2 Practice with Feedback

After reviewing the Rule Card, the learner is presented with real-world practice items, which are representative of the environment under which the learner will apply their knowledge (Green and Seifert, 2005). The practice items are a series of inputs that might be provided to the AI/ML system. In this hands-on experience, optimally, the learner compares the novel situation to the rules they have been taught, and the learner is asked to respond with their prediction of the system's output based on the application of the proper rule. After each practice item, if the learner responds correctly, this is reinforced with a statement telling the learner that they are correct, followed by a reiteration of the rule's application to the novel situation. If the learner responds incorrectly, they are told that they were incorrect, which rule should have been applied, and why the appropriate rule was applicable (Figure 2.2).

Figure 2.2

Sample Practice with Feedback Item



This step is important, as it provides a feedback loop, continuously evaluating and improving the learner's performance. This safe and supportive environment allows the learner to receive immediate information on their accuracy, and they are given the opportunity to adjust their behavior and strategies in future practice and test items, with the aim of improving their skills (VanLehn, 2011). It also increases the learner's motivation and engagement due to the visible progress.

An exemplar-based training program might just as easily provide feedback; however, with an inference-based knowledge base, and after many inferences have been made based on a large quantity of exemplars, it's more difficult to determine exactly where the learner made the error. This would make it difficult to repair their knowledge (VanLehn, 2011).

The Practice with Feedback phase also helps with the skill acquisition. A theory proposed by Fitts and Posner (1967), stated that the learner progresses through three phases of skill acquisition: cognitive (the mental processes of learning and understanding concepts), associative (translating declarative knowledge into procedural knowledge), and finally autonomous (when the learned skill is almost automatic and requires minimal thought).

Applying this theory to rule-based training, in the cognitive phase, learners study and memorize the rules, factual and counterfactual exemplars, and principles (similar to Kahneman's System Two thinking). In the associative phase, the learners using a rule-based training program would practice their skill, applying the rules they have learned, refining their technique and becoming more efficient as they apply the rules more consistently and accurately. Finally, the autonomous phase would involve learners applying the rules automatically with less conscious thought or effort (Kahneman's System One thinking), attaining a higher level of accuracy. The goal of the Practice with Feedback phase is to provide this highly visible, upwardly mobile proficiency level to the learner.

2.5.3.3 Test

In our research, after the training (Rule Card presentation and Practice with Feedback), the learner is given real-world test items. This is done to measure the effectiveness of the training. Although not an integral part of Explicit Rule Learning, this phase may be included not only to measure the learners' proficiency, but also to evaluate the effectiveness and comprehensiveness of the rules. Assuming the test represents the realm

of real-world situations the learners might encounter, to the degree possible, the rules would optimally also represent the breadth of the real-world situations.

The responses to the test items might be in the form of multiple choice, true/false, or open-ended (long- or short- answer) responses. The scoring of these items might include confidence ratings (Figure 2.3), or simply correct/incorrect (Figure 2.4).

Figure2.3

Test Item With Confidence Ratings

For example, a participant might be asked to predict a system's output, and the multiple-choice responses (with confidence ratings) might be:

a) the system will definitely output x

b) the system will likely output x

c) I'm unsure of what the system's output will be

d) the system will likely output y

e) the system will definitely output y

Figure 2.4

Test Item With Correct/Incorrect Scoring (Without Confidence Ratings)

This might also be asked without confidence levels:

a) the system will output x

b) the system will output y

2.5.4 Explicit Rule Learning Content

The content of the Explicit Rule Learning tutorials (i.e., the rules that are selected) varies with the AI/ML system on which the learner is being trained. However, the following basic properties should be considered so the learner develops a better mental model of the system, and gains a deeper understanding of the underlying principles, rationale, logic, and the weight the system tends to assign variables:

- The overall goal of ERL is to accelerate the proficiency of the human learning the AI/ML system, and to garner higher trust levels and more accurate mental models of the system
- Content can be developed by a novice of the system, within 6 weeks of using the AI/ML system
- Rules are global explanations
- Content is based on expert feedback, patterns observed in the system
- Rules that have a higher probability of occurring should take precedence over those that are not
- Rules should be given priority if they are more severe (e.g., more dangerous in a driving task), or have harsher consequences
- Rules are clear, concise, and have boundaries (i.e., to the extent possible, there should not be any ambiguous rules that might overlap with another rule's logic)
- The input variables and their possible values are described; the output variables, their type, and possible values are described; the boundaries are described to the learner as there are almost always degrees of membership (Hampton, 1998)
- The reasoning/logic behind the rule are explained to the extent possible

- Rules contain the context in which they are to be applied
- Any known limitations or uncertainties are described
- Rules contain information on the circumstances they are to be used, and the possible differences in the system's output based on various circumstances/scenarios

Explicit Rule Learning was developed and tested throughout the progression of five studies. We began by training participants in an ML image classifier, using exemplars. This transitioned into comparing exemplar-based training with rule-based training. The results of these studies led to the inception of Explicit Rule Learning, and informed its framework and structure. Finally, Explicit Rule Learning was tested on a real-world, more advanced AI/ML system. Next, the five studies will be described.

3 Explicit Rule Learning Studies: General Method for Studies 1-4 (MNIST Studies)

Although the AI/ML systems and training method/presentation varied between studies, a standard protocol was followed for all of the studies. Studies 1-4 trained learners on an image classifier, and Study 5 trained learners in a full self-driving autonomous vehicle AI system.

In all of these within-subject studies, participants received training and were asked to predict the output of an intelligent software system. Some of the training was exemplar-based, some rule-based, and sometimes the participants did not receive training (control conditions). After the training, participants were given test scenarios and asked to predict the output of the AI/ML system. The predictions were scored for accuracy and compared across the different training methods and presentations.

Further details on the variations between studies will be detailed separately. Following is the standard protocol followed for all of the studies.

3.1 Participants

Participants were undergraduate students from Michigan Technological University who participated in the online studies in exchange for Introduction to Psychology course credit. The participants were novices in the domains (i.e., they had no experience with the AI/ML systems used in the studies).

3.2 Training and Testing Protocol

Participants received training that was exemplar-based, rule-based, or a combination thereof; there were also control conditions where the participants did not receive training. The stimuli for each study were subdivided into four categories. For example, Studies 1-4 used four pairings of digits as stimuli (0 and 6, 1 and 5, 2 and 3, 4 and 9) and four training conditions. Study 5 had four issues with which the Tesla FSD had problems (stop signs, turning, driving in the wrong lane, driving in the center) and two training conditions (Explicit Rule Learning and no training). The participants received training on a subset of stimuli. Then, all participants were tested on all of the stimuli. Therefore, the test items upon which participants were trained were the experimental conditions, and the test items upon which participants were untrained were the control conditions.

In all of the studies the test consisted of some input given to the AI/ML system, and the participants' task was to predict the output of the AI/ML system. The input contained objects and features that, when considered as a whole, would render a likely output by the system.

3.3 Procedure

Participants completed this study online via Qualtrics Survey platform. Participants were randomly assigned to one of several randomly assigned versions of the study (four versions for Studies 1-4, and two versions for Study 5). After consenting to participate in the study, participants responded to demographic (Studies 1-5) and various other

questionnaires (Study 5), followed by training. Next, participants were given a test, where they were asked to predict the output of the AI/ML system given a specific input.

3.4 Coding Scheme

The predictions made by the participants in the test cases were scored for accuracy. The responses to prompts in the questionnaires were coded (some quantitatively, some qualitative).

3.5 Analysis

Although the coding varied slightly between studies, the general coding scheme was as follows. The mean accuracy in predicting test cases was calculated, with comparisons made between the different training methods/presentations and the control condition. An ANOVA was performed to identify any statistical differences between the different training methods/presentations and the different stimuli groupings (digit pairings for Studies 1-4, rules for Study 5). The Tukey Test was run to identify whether there were any statistical differences between the means of the results of the training methods/presentations. Lastly, the results of the questionnaires were summarized.

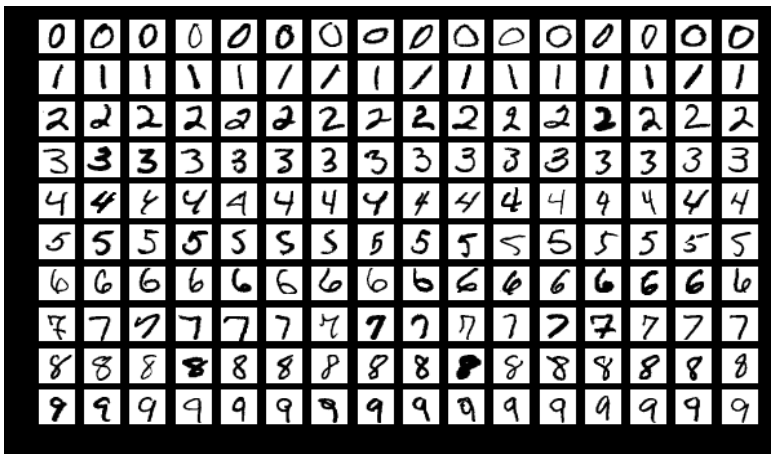
Next, each of the study methods and results will be described individually.

4 MNIST Study 1

The stimuli for Studies 1-4 came from the Modified National Institute of Standards and Technology (MNIST) database (Figure 4.1). This is a large database of individually written handwritten digits that is commonly used for training various image classification systems. Classifiers are "trained" in the ten digits (0-9) with a subset of digits. After the training, the classifier is given a new set of unseen digits called "test" items and tasked with classifying these. For example, after the classifier has been trained, it is shown a "3" it hasn't seen before and asked to classify it. It might say it's a "3", but it also might say it's an "8", or a "4", or some other digit.

Figure 4.1

Samples of Digits from the MNIST Database



We used the Discovery Platform (http://obereed.net.3838/mnist), which uses a linear SVM (support vector machine) classifier trained on 10,000 cases (5,000 correctly classified). Simply put, this type of machine learning algorithm identifies decision boundaries for different classifications (i.e., classifications of "1" or "3", etc.). If the

stimuli it is provided with fits into a specific boundary, it is classified as a member of that class. If it does not fit into that class, it is classified as a member of another class.

The Discovery Platform achieved an 89.2% accuracy rate with test items. However, for our studies, the participants were told that the classifier accurately classified the digits 50% of the time, and incorrectly classified the digits 50% of the time. Therefore, stimuli cases were hand-selected to conform to the .50 base rate, and participants were asked to assume the training and practice material, as well as test items were classified correctly 50% of the time.

The research questions for Study 1 were:

R1: Are participants more accurate in predicting the output of the MNIST image classifier after receiving Positive (factual exemplars) training versus Negative (counterfactual exemplars) training?

R2: Does presenting both Positive and Negative exemplars concurrently improve the participants' prediction accuracy?

R3: Are participants more accurate in predicting the output of the MNIST image classifier after receiving training presented to them in an interleaved versus a blocked presentation?

4.1 Method

4.1.1 Participants

58 college-aged students (64% male), average age of 20 years (SD = 1.06) completed the within-subject 45-minute study in exchange for Introduction to Psychology course credit.

4.1.2 Classifier Prediction Task

We systematically selected 4 “digit pairings” to use as stimuli (Table 4.1). Stimuli pairings were selected based on a confusion matrix of the SVM classifier’s image classes and labels, where positive and negative predicted classes of specific pairings were similarly aligned in space locations. Also, a correspondence map was made on the positive and negative predicted classes, where the pairings we selected demonstrated similar relative relationships.

Table 4.1

Study 1 Digit Pairings

Positive cases (correctly classified)	Negative Cases (incorrectly classified)
0 classified as **0** **6** classified as **6**	**0** classified as **6** **6** classified as **0**
1 classified as **1** **5** classified as **5**	**1** classified as **5** **5** classified as **1**
2 classified as **2** **3** classified as **3**	**2** classified as **3** **3** classified as **2**
4 classified as **4** **9** classified as **9**	**4** classified as **9** **9** classified as **4**

Additionally, the exemplar-based training showed examples of positively and/or negatively classified digits. These were shown to the participants either in an interleaved or blocked presentation, which was based on a study done by Rau et al., (2010), who found differences in students' performance after receiving interleaved vs blocked training.

We used a Graeco-Latin Square to counterbalance four versions/groups, balancing the training method and digit pairings (Table 4.2).

Table 4.2

Study 1 Graeco-Latin Square with Digit Pairings and Training Conditions

Version	Condition/ Pairing 1	Condition/ Pairing 2	Condition/ Pairing 3	Condition/ Pairing 4
1	Positive Interleaved (1-5)	Positive Negative Interleaved (0-6)	Negative Interleaved (4-9)	Positive Negative Blocked (2-3)
2	Positive Negative Interleaved (2-3)	Positive Interleaved (4-9)	Positive Negative Blocked (0-6)	Negative Interleaved (1-5)
3	Negative Interleaved (0-6)	Positive Negative Blocked (1-5)	Positive Interleaved (2-3)	Positive Negative Interleaved (4-9)
4	Positive Negative Blocked (4-9)	Negative Interleaved (2-3)	Positive Negative Interleaved (1-5)	Positive Interleaved (0-6)

Participants completed the study at the location of their choice on www.Qualtrics.com. They were given the following instructions:

> "This task involves an AI system that classifies digits. For example, the system is shown the digit 2 and it must return a label in the form of a digit. It might say it's a 2, or it might make an error and say it's a 1 or a 3 or some other digit.
>
> Your job will be to tell me how the system will classify the digit. Regardless of whether or not you recognize the digit, your response should always be what you think the system will classify it as.

> Next, you will be shown some examples to familiarize yourself with the system and how it works. Your job is to learn where the system succeeds and where it fails. Then following the training, you'll be tested on cases where you don't know the answer. You'll go through four sections like this, where first there's training, and then there's a test. At the end of the four sections, you'll be asked to respond to a short demographic survey."

Next, each condition/digit pairing in the Graeco-Latin square displayed 40 training items. For the Blocked Presentation, participants were shown 4 screens with 10 examples each. For the Interleaved Presentation, participants were shown 8 screens with 5 examples each.

Each digit pairing training section was followed immediately with 32 test items for that digit pairing. Again, the test items were configured so that 50% were correctly classified. When responding to each test item, participants conveyed their confidence level for their response; the choices were: "Definitely classified correctly", "Probably classified correctly", "I don't know", "Probably classified incorrectly", or "Definitely classified incorrectly".

4.1.3 Demographic Questionnaire

Lastly, participants were asked to respond to a demographic questionnaire that asked for their age and gender, to determine the representativeness of the sample and its ability to be generalized to a broader population.

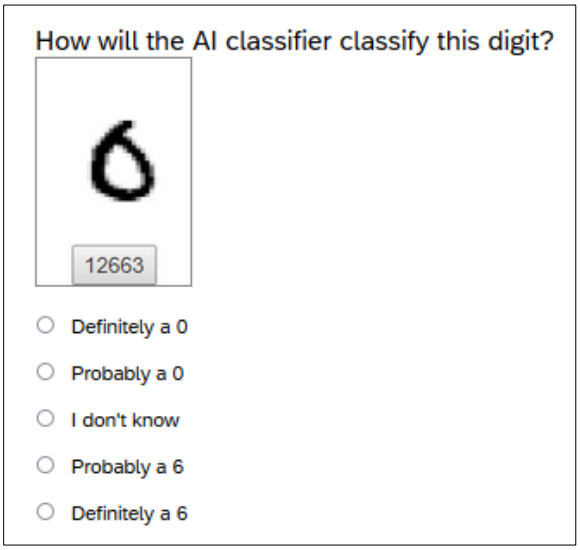
4.1.4 Coding Scheme

Participants completed a total of 128 test items, 32 for each digit pairing. Figure 4.2 displays a sample test item. In this example, the classifier was given a "0" as input, but it misclassified it as a "6". If the participant responded with "Definitely a 6" or "Probably a 6", they were given 1 point for accuracy. If they responded with any of the other choices, they were given 0 points. These accuracy points were used to analyze the results.

Initially, accuracy was scored with consideration to participants' confidence levels with each item having a possible 1 (lowest accuracy) to 5 (highest accuracy) points. For example, using the example in the previous paragraph, participants received 5 points if they responded with "Definitely a 6", 4 points if they responded with "Probably a 6", 3 points for "I don't know", 2 points for "Probably a 0", and 1 point for "Definitely a 0". The two modes of scoring (confidence ratings 1-5 points versus correct/incorrect) were compared and the difference was found to be negligible. Therefore, for the sake of simplicity, in this study as well as Studies 2-4, accuracy was scored as correct/incorrect, without consideration to confidence levels.

Figure 4.2

Sample Test Item from Study 1



4.2 Results

An ANOVA (Type II Wald $\chi 2$ tests) revealed that there was a statistically significant difference in the different training types ($\chi 2$ (15.12) = 3, $p < .05$), and the different digit pairings, ($\chi 2$ (16.09) = 3, $p < .05$). The Tukey Test revealed statistically significant differences when comparing Positive Interleaved with Negative Interleaved Training and comparing Pos/Neg Interleaved and Blocked Training to Positive Interleaved Training.

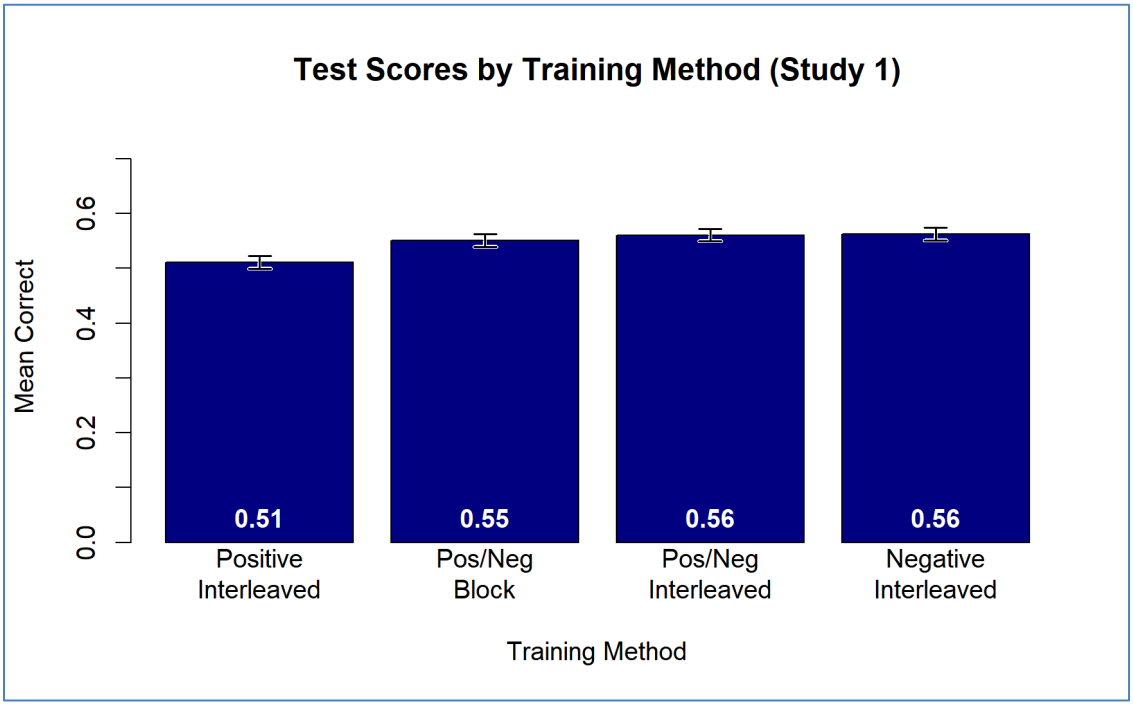
The accuracy of the participants' predictions was collapsed across digit pairings and compared across training methods.

R1: Are participants more accurate in predicting the output of the MNIST image classifier after receiving Positive (factual exemplars) training versus Negative (counterfactual exemplars) training?

As shown in Figure 4.3, participants' accuracy was just barely above chance for all of the conditions. However, there was slight improvement in cases where participants' training contained negative examples (Positive/Negative Blocked presentation, Positive/Negative Interleaved presentation, and Negative Interleaved).

Figure 4.3

Study 1 Accuracy Results by Training Method. Error bars represent standard error.



R2: Does presenting both Positive and Negative exemplars concurrently improve the participants' prediction accuracy?

Based on the results, it is clear that showing negative exemplars improved the participants' ability to accurately predict the system's classification. However, it does not

seem that the combination of Positive + Negative made a difference versus Positive exemplars alone, at least not as much as the exposure to Negative exemplars.

R3: Are participants more accurate in predicting the output of the MNIST image classifier after receiving training presented to them in an interleaved versus a blocked presentation?

The results show that there was not a significant difference between the Positive/Negative Interleaved and Positive/Negative Block presentation. Additionally, the accuracy scores were about the same, with a one percent higher accuracy for the interleaved presentation.

4.3 Discussion

We ran a post hoc power analysis to determine whether or not the sample size was sufficient. We found that we needed n >= 24 to find our effect, so our sample of n = 58 was sufficient.

In this study, participants were shown exemplars of both correctly (positive) and incorrectly (negative) classified digits. Interestingly, although in general the participants scored barely above chance, when the participants were trained with negative counterfactual examples, they performed slightly better. Aligned with previous findings, showing errors (Cattaneo and Boldrini, 2017; van der Meij and Flacke, 2020) and counterfactual cases that defined boundaries (Kuhl et al., 2023) helped the most and were most critical.

Any schemas the participants inferred from the examples on the rationale used by the classifier when classifying the digits were based on their intuition and their interpretation of the examples. We are unable to identify what criteria the participants were relying on to predict the classifications. Perhaps they found a pattern, or similarity, or somehow formed mental categories along with rules of membership that determine the category in which the stimuli belong. Maybe the participants anthropomorphized the classifier and used their human "expertise" in identifying factual and counterfactual reasons a digit might be correctly or incorrectly classified. Without further probing, we are unable to determine the participants' reasoning, logic, or rationale for their predictions. However, it would be reasonable to assume that the participants were inferring some type of conglomerate logic put together by patterns of features they perceived in the exemplars.

Chi et al., (1989) described learners' self-explanation abilities to create inference rules from examples, whereby learners used these rules to form instantiations and definitions that could be used to generalize to new situations. These inferred rules are more operational than the exemplars alone, complete with applicable conditions, converting declarative instances into usable procedures.

Study 2 further explored rule-based training. Specifically, the goal was to present the participants with probabilistic, verbalizable rules, based on an expert evaluation of the stimuli, rather than the inferred rules naïve learners created on their own, based on exemplars. These rules would be complemented with factual and counterfactual

exemplars. To accomplish this, we created a rule-based Cognitive Tutorial. Our hope was that the rules would be sensitive enough to predict classifications better than chance (>.50), with the aim of improving the participants' accuracy when predicting the system's classification on test items.

Additionally, for Study 2, we addressed a limitation from Study 1: the digit pairings were bi-directional. For example, we used 1s that were classified as 1, 1s that were classified as 5, 5s that were classified as 5, and 5s that were classified as 1. Also, although the training and test stimuli were 50% correct and 50% incorrect, the stimuli were randomly selected, with no regard for patterns or features. For the next study, stimuli were selected so that the digit pairings were unidirectional: 1s classified as 1, and 1s classified as 5 (omitting 5s classified as 5, and 5s classified as 1). Also, stimuli were systematically chosen following a detailed analysis of patterns and features.

5 MNIST Study 2

Study 2 followed the same method and protocol as Study 1. However, the digit pairings were unidirectional (Table 5.1), and the conditions were changed. We kept the Positive + Negative Interleaved condition but removed the other three conditions from Study 2. In their place, we added a control condition (No training), and two Explicit Rule Learning conditions: one with the Rule Card + Practice with Feedback, and one without the Rule Card, only presenting the participant with the Practice with Feedback component.

Table 5.1

Study 2 Digit Pairings

Positive cases (correctly classified)	Negative Cases (incorrectly classified)
0 classified as **0** **1** classified as **1** **3** classified as **3** **4** classified as **4**	**0** classified as **6** **1** classified as **5** **3** classified as **2** **4** classified as **9**

Rule Content

In order to decide the content of the rules, the stimuli were manually inventoried, with the goal of identifying fact-based patterns and tendencies made by the classifier. For example, when comparing scores of 4s classified as 4 with 4s classified as 9 side-by-side, it was clear that “closed-top” 4s were often misclassified as 9 (Figure A.7). For each digit

pairing, two to five such patterns were identified. Next, a base rate of occurrence was determined.

In order to determine the true base rate, 200 cases were evaluated for each observed pattern. For example, in the case mentioned in the previous paragraph, closed-top 4s were misclassified as 9. To identify the true base rate, 100 unique 4s with closed-tops were identified from the MNIST database. The number of cases where these closed-top 4s were classified as 9 was tallied, as was the number of cases where closed-top 4s were classified as 4. Conversely, 100 cases where the 4s *did not have* closed-tops were also identified from the MNIST database. The number of cases where these "open-top" 4s were classified as 9 was tallied, as was the number of cases where these open-top 4s were classified as 4.

This resulted in a list of base rates for approximately twenty patterns (two to five patterns for each of the four digit pairings). Lastly, the pattern base rates were compared for each digit pairing, and the two most frequently occurring patterns (i.e., the two highest base rates) were selected as the content for the rules that would be created for the training. This resulted in two rules for each of the digit pairings, for a total of eight rules.

Finally, factual and counterfactual exemplars were selected from the MNIST database for each rule. These would be presented to the participant on the Rule Card as a visual representation of the rule, and as training and test items. These were also used as training stimuli in the exemplar-based training condition.

For Study 2, the wording in the instructions given to the participants was updated, reiterating that they should be predicting how the system will classify the digit rather than the participant responding with how they, as a human well-versed in reading digits, would classify the digit. The stimuli from Study 1 were reviewed, and random stimuli that did not conform to the exemplars or rules were taken out, and systematically replaced with more relevant cases.

The research questions for Study 2 were:

R1: Are participants more accurate in predicting the output of the MNIST image classifier after receiving Positive/Negative Interleaved training versus Explicit Rule Learning training?

R2: Does presenting both components of Explicit Rule Learning (Rule Card + Practice with Feedback) improve participants' prediction accuracy as compared to Explicit Rule Learning (Practice with Feedback only)?

5.1 Method

5.1.1 Participants

51 college-aged students (53% male), average age of 20 years (SD = 2.50) completed the within-subject 45-minute study in exchange for Introduction to Psychology course credit.

5.1.2 Classifier Prediction Task

Participants completed the study at the location of their choice on www.Qualtrics.com. We counterbalanced the conditions and digit pairings into four versions of the study using a Graeco-Latin square (Table 5.2).

Table 5.2

Study 2 Graeco-Latin Square with Digit Pairings and Training Conditions

Version	Condition/ Pairing 1	Condition/ Pairing 2	Condition/ Pairing 3	Condition/ Pairing 4
1	Explicit Rule Learning: Practice with Feedback Only (1-5)	Positive Negative Interleaved (0-6)	No Training (4-9)	Explicit Rule Learning: Rule Card + Practice with Feedback (3-2)
2	Positive Negative Interleaved (3-2)	Explicit Rule Learning: Practice with Feedback Only (4-9)	Explicit Rule Learning: Rule Card + Practice with Feedback (0-6)	No Training (1-5)
3	No Training (0-6)	Explicit Rule Learning: Rule Card + Practice with Feedback (1-5)	Explicit Rule Learning: Practice with Feedback Only (3-2)	Positive Negative Interleaved (4-9)
4	Explicit Rule Learning: Rule Card + Practice with Feedback (4-9)	No Training (3-2)	Positive Negative Interleaved (1-5)	Explicit Rule Learning: Practice with Feedback Only (0-6)

Consideration was given to the fact that the Rule Card (Figure 2.1) screens contain more content than the exemplar screens. In order to ensure that the participants carefully read and considered each rule, a one-minute timer was put on the screens where the Rule Card

was being displayed. Participants were told that they should carefully review the Rule Card, and that they would be able to proceed (i.e., click on the "Next" button) after one minute. The participants were able to spend as much time reviewing the Rule Card as they needed; the timer only ensured the minimum amount of time they had to view the Rule Card.

Participants were given the following instructions:

> "This task involves an AI system that classifies digits. For example, the system is shown the digit 2 and it must classify it in the form of a digit. It might say it's a 2, or it might make an error and say it's a 1 or a 3 or some other digit.
>
> Your job will be to tell me how the system will classify the digit. Regardless of whether or not you recognize the digit, your response should always be what you think the system will classify it as.
>
> - In some cases, you will be shown some examples to familiarize yourself with the system and how it works. Your job is to learn where the system succeeds and where it fails. Then following the training, you'll be tested on cases where you don't know the answer.
> - In other cases, there won't be any training, and you'll be asked to predict how the system works without any training.

- At the end of the four sections, you'll be asked to respond to a short demographic survey."

5.1.3 Demographic Questionnaire

Lastly, participants were asked to respond to a demographic questionnaire that asked for their age and gender, to determine the representativeness of the sample and its ability to be generalized to a broader population.

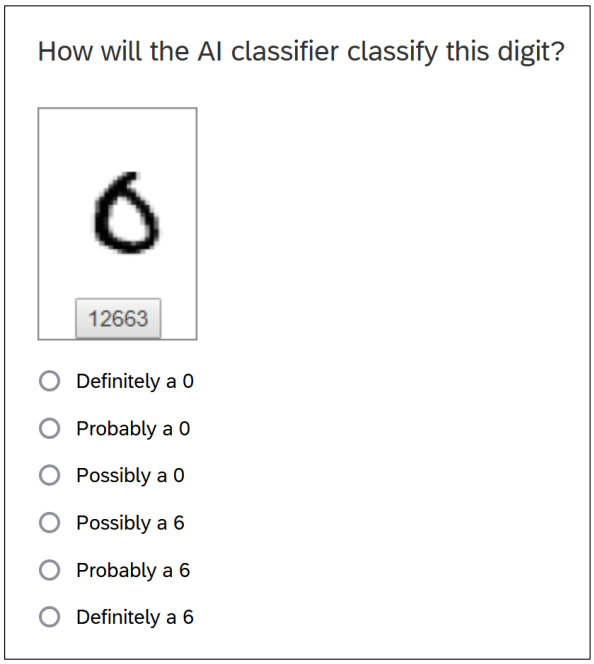
5.1.4 Coding Scheme

Participants completed a total of 128 test items, 32 for each digit pairing. Participants were required to respond with their prediction of the system's classification using a 6-point confidence rating. In comparison to Study 1, which used a 5-point scale and allowed participants to make a neutral selection, in Study 2, participants selected from a 6-point scale: "Definitely correct", "Probably correct", "Possibly correct", "Possibly incorrect", "Probably incorrect", and "Definitely incorrect". In other words, we eliminated the "I don't know" choice, and the participants were required to make a correct/incorrect prediction decision.

Figure 5.1 displays a sample test item. In this example, the classifier was given a "0" as input, but it classified it as a "6". If the participant responded with "Definitely a 6", "Probably a 6" or "Possibly a 6", they were given 1 point for accuracy. If they responded with any of the other choices, they were given 0 points. These accuracy points were used to analyze the results.

Figure 5.1

Sample Test Item for Study 2



5.2 Results

An ANOVA (Type II Wald $\chi 2$ tests) revealed that there was a statistically significant difference in the different training types ($\chi 2$ (47.3) = 3, $p < .05$), and the different digit pairings, ($\chi 2$ (150.17) = 3, $p < .05$). The Tukey Test revealed statistically significant differences when comparing each of the Training types with one another.

The accuracy of the participants' predictions was collapsed across digit pairings and compared across training methods.

R1: Are participants more accurate in predicting the output of the MNIST image classifier after receiving Positive/Negative Interleaved training versus Explicit Rule Learning training?

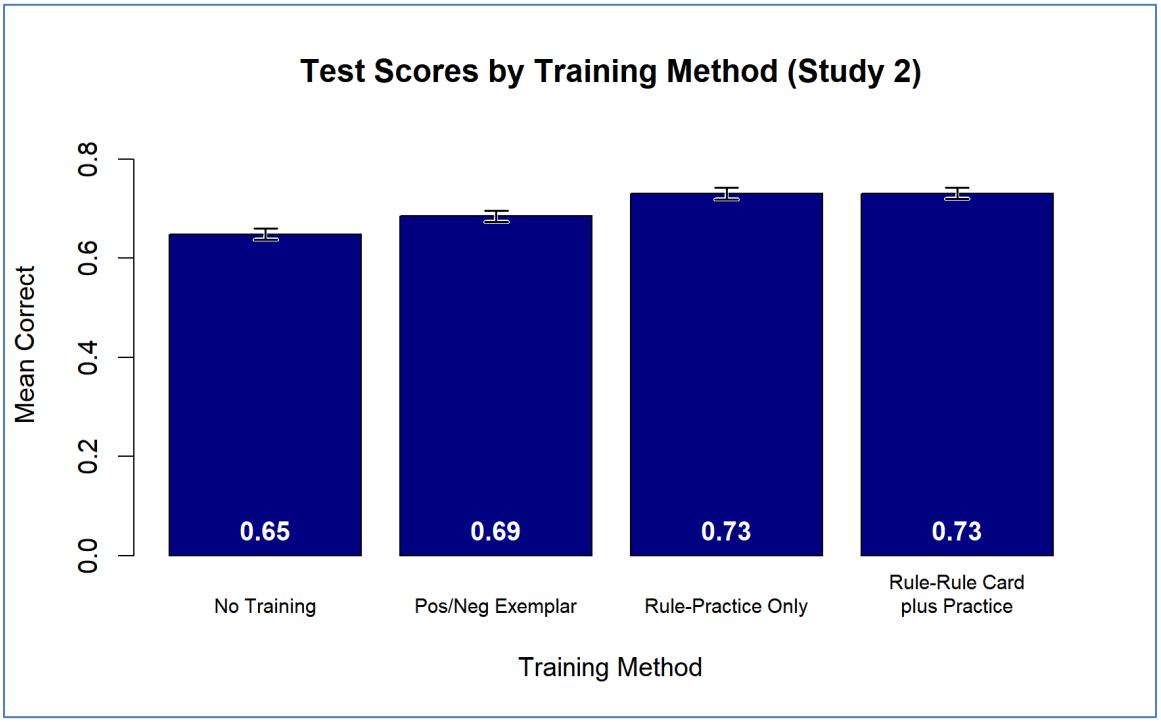
R2: Does presenting both components of Explicit Rule Learning (Rule Card + Practice with Feedback) improve participants' prediction accuracy as compared to Explicit Rule Learning (Practice with Feedback only)?

Figure 5.2 shows that participants did better in all conditions, when compared to Study 1. Being more systematic about the stimuli is believed to have made this difference.

Both Explicit Rule Learning conditions had higher scores than exemplars alone.

Figure 5.2

Study 2 Accuracy Results by Training Method. Error bars represent standard error.



5.3 Discussion

This study used No Training as a control condition, and also compared Exemplar Training with Explicit Rule Learning. The explicit, probabilistic, verbalizable rule guided the participants towards specific "if…then" causal patterns. The stimuli were unidirectional. Additionally, we changed the wording in the instructions, clarifying that their role was to predict if the classifier would be correct, rather than using their own intuition in differentiating the digits. In other words, a human might clearly see how a zero might be mistaken for a six, but the participants were asked to think of the

classifier's output, and not use their own judgment. These modifications seem to have improved participants' accuracy in predicting the classifier's output.

We also teased apart the Explicit Rule Learning training – Rule Card + Practice with Feedback vs Practice with Feedback alone. It seems that the Rule Card didn't matter. As long as the participants received feedback, this is what helped them to be more proficient in predicting the classifier's output. This was investigated further in the next study. Perhaps the Rule Card might not be useful during the initial training and could be more useful as a reference card later on when using the system, but further research isolating the Rule Card only (without Practice with Feedback) was the next step.

In comparing exemplar-based training with explicit rule-based training, exposing the participant to factual and counterfactual exemplars alone was not as effective as probabilistic rule-based training. This finding supports the usage of Cognitive Tutorials (Mueller et al., 2021), which utilizes global explanations. The learners who applied global reasoning to their predictions fared better compared to learners who were shown examples (local explanations), and who had limited ability to forecast future cases. After all, the learner should understand how the system works in general (Wick and Thompson, 1992) and be able to apply their global understanding of the system to future, unseen cases, thereby utilizing effective skill transference.

In the next study, we eliminated more of the randomness of the stimuli. Each stimulus was specifically selected with application to one rule and only that rule, removing the ambiguity of possible applications to more than one rule.

6 MNIST Study 3

The conditions and Digit Pairings for Study3 were the same as those for Study 2. However, the stimuli for Study 3 were more curated.

First, we identified instances of stimuli that were duplicated in the study. For example, stimuli appeared on the Rule Card, and then again as a test item. The duplicates were removed, and all of the stimuli used in the Explicit Rule Learning training and test items were unique. There was one exception to this policy. If the stimuli used in the Explicit Rule Learning conditions were different from those used in the exemplar-based condition, it might be reasonable to wonder whether any effect we find was due to the training condition or if it was due to the different stimuli presented to the participant. For this reason, the exemplars on the Rule Cards were the same as the exemplars in the Positive-Negative Interleaved exemplar-based condition. In other words, a participant that was assigned to version 1 of the study, who is presented with Explicit Rule Learning training for the digit pairing of 3s and 2s saw the same training stimuli as another participant who was presented with exemplar-based Positive-Negative Interleaved training for the digit pairing of 3s and 2s.

Secondly, we identified instances where some of the stimuli were ambiguous, with possible application to more than one rule, in some cases, with conflicting results. In this study, we carefully selected stimuli that could only be applied to one rule.

6.1 Method

6.1.1 Participants

Forty-five college-aged students (69% male), average age of 20 years (SD = 1.70) completed the within-subject 45-minute study in exchange for Introduction to Psychology course credit.

6.1.2 Classifier Prediction Task

Participants completed the study at the location of their choice on www.Qualtrics.com. The conditions and digit pairings were counterbalanced into four versions using the same Graeco-Latin square used for Study 2. (Table 5.2).

Again, as in Study 2, a timer was instantiated on the screens with the Rule Card, to ensure that the participants would spend at least one minute reviewing the content.

Participants were given the following instructions:

> "This task involves an AI system that classifies digits. In this project, a human is asked to draw, for example, a 5. Then the AI system is shown a digital copy of the 5 the human drew. The AI system is asked to classify the 5. The AI system might say it's a 5, or it might make an error and say it's a 1 or a 3 or some other digit.
>
> Your job will be to tell me how the system will classify the digit. Regardless of whether or not you recognize the digit, your response should always be what you think the system will classify it as.

- In some cases, you will be shown some examples to familiarize yourself with the system and how it works. Your job is to learn where the system succeeds and where it fails. Then following the training, you'll be tested on cases where you don't know the answer.
- In other cases, there won't be any training, and you'll be asked to predict how the system works without any training.
- At the end of the four sections, you'll be asked to respond to a short demographic survey."

6.1.3 Demographic Questionnaire

Lastly, participants were asked to respond to a demographic questionnaire that asked for their age and gender, to determine the representativeness of the sample and its ability to be generalized to a broader population.

6.1.4 Coding Scheme

The coding scheme for Study 3 was the same as Study 2. Participants completed a total of 128 test items, 32 for each digit pairing.

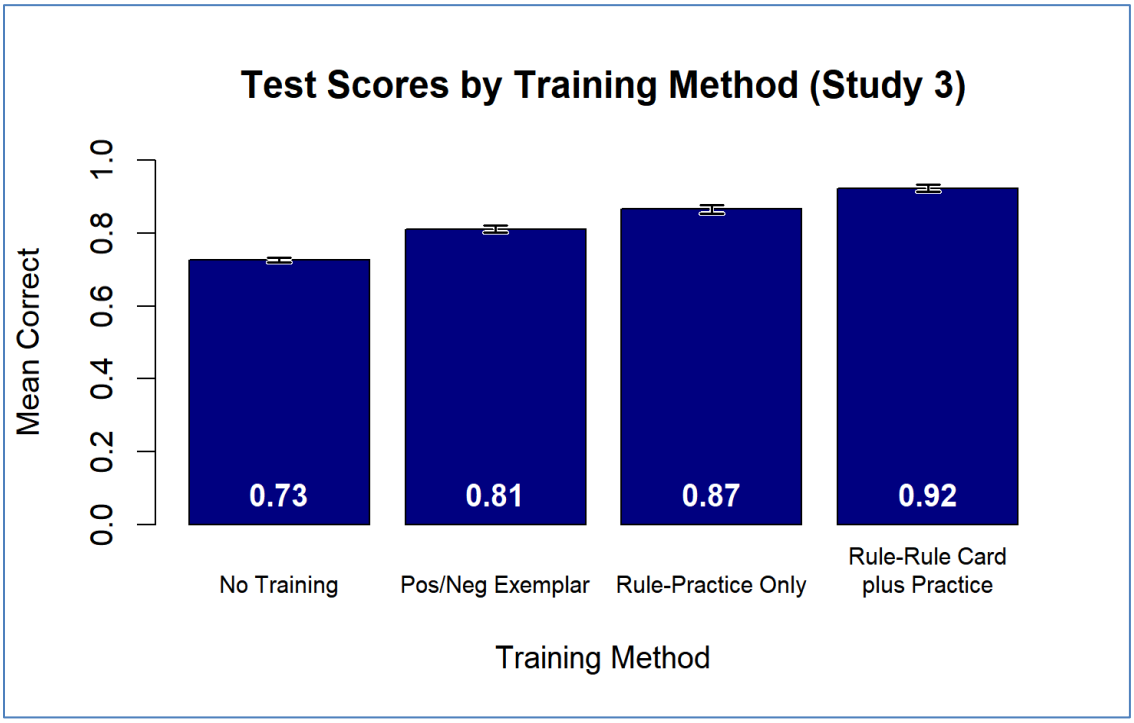
6.2 Results

An ANOVA (Type II Wald $\chi 2$ tests) revealed a statistically significant difference in the different training types ($\chi 2$ (231.03) = 3, $p < .05$), and the different digit pairings, ($\chi 2$ (101.46) = 3, $p < .05$). The Tukey Test revealed statistically significant differences when comparing each of the Training types with one other. Participants had better prediction accuracy scores in all conditions versus Study 2 (Figure 6.1). Being more systematic by eliminating duplicate stimuli (the same digit image appearing in training and as a test

item) and being intentional about the stimuli in the test (ensuring each stimulus was an application of one rule, and one rule only), seems to have given us a clearer picture of the difference between Exemplar vs Explicit Rule Learning Training, with Explicit Rule Learning providing a more effective training method.

Figure 6.1

Study 3 Accuracy Results by Training Method. Error bars represent standard error.



6.3 Discussion

This study used No Training as a control condition, and also compared Exemplar Training with Explicit Rule Learning, just as Study 2 did. Being systematic and intentional with the stimuli selection appears to have made a difference, in that the participants were better predictors of future cases.

Our last study in this series compared the various components of Explicit Rule Learning: Rule Card only, Practice with Feedback only, Rule Card + Practice with Feedback.

7 MNIST Study 4

The only update to Study 4 from Study 3 was the condition comparison. The complete Explicit Rule Learning method consists of a Rule Card, followed by Practice with Feedback. The intent was to present the participant with a probabilistic, verbalizable rule, complete with factual and counterfactual exemplars, followed by practice problems. It is hoped that this comprehensive, multifaceted approach will provide the participant with better representation, a more complete and robust mental model, and a stronger ability to generalize their knowledge to future, unseen situations. In the final study of this series, we tested each of the two components of Explicit Rule Learning (Rule Card + Practice with Feedback) against the complete form and compared these three to a No Training condition (control condition). The Digit Pairings were the same as those used in Studies 2 and 3.

The research questions for Study 4 were:

R1: Are participants more accurate in predicting the output of the MNIST image classifier after receiving both components of Explicit Rule Learning (Rule Card + Practice with Feedback), Explicit Rule Learning (Rule Card only), or Explicit Rule Learning (Practice with Feedback only) training?

7.1 Method

7.1.1 Participants

Forty-six college-aged students (39% male), average age of 19 years (SD = 1.01) completed the within-subject 45-minute study in exchange for Introduction to Psychology course credit.

7.1.2 Classifier Prediction Task

Participants completed the study at the location of their choice on www.Qualtrics.com. We counterbalanced the conditions and digit pairings into four versions of the study using a Graeco-Latin square (Table 7.1).

Table 7.1

Study 4 Graeco-Latin Square with Digit Pairings and Training Conditions

Version	Condition/ Pairing 1	Condition/ Pairing 2	Condition/ Pairing 3	Condition/ Pairing 4
1	Explicit Rule Learning: Practice with Feedback Only (1-5)	Explicit Rule Learning: Rule Card Only (0-6)	No Training (4-9)	Explicit Rule Learning: Rule Card + Practice with Feedback (3-2)
2	Explicit Rule Learning: Rule Card Only (3-2)	Explicit Rule Learning: Practice with Feedback Only (4-9)	Explicit Rule Learning: Rule Card + Practice with Feedback (0-6)	No Training (1-5)
3	No Training (0-6)	Explicit Rule Learning: Rule Card + Practice with Feedback (1-5)	Explicit Rule Learning: Practice with Feedback Only (3-2)	Explicit Rule Learning: Rule Card Only (4-9)
4	Explicit Rule Learning: Rule Card + Practice with Feedback (4-9)	No Training (3-2)	Explicit Rule Learning: Rule Card Only (1-5)	Explicit Rule Learning: Practice with Feedback Only (0-6)

Again, as in Studies 2 and 3, a timer was instantiated on the screens with the Rule Card, to ensure that the participants would spend at least one minute reviewing the content. Participants were given the same instructions that were given in Study 3:

> "This task involves an AI system that classifies digits. In this project, a human is asked to draw, for example, a 5. Then the AI system is shown a digital copy of the 5 the human drew. The AI system is asked to classify the 5. The AI system might say it's a 5, or it might make an error and say it's a 1 or a 3 or some other digit.

Your job will be to tell me how the system will classify the digit. Regardless of whether or not you recognize the digit, your response should always be what you think the system will classify it as.

- In some cases, you will be shown some examples to familiarize yourself with the system and how it works. Your job is to learn where the system succeeds and where it fails. Then following the training, you'll be tested on cases where you don't know the answer.
- In other cases, there won't be any training, and you'll be asked to predict how the system works without any training.
- At the end of the four sections, you'll be asked to respond to a short demographic survey."

7.1.3 Demographic Questionnaire

Lastly, participants were asked to respond to a demographic questionnaire that asked for their age and gender, to determine the representativeness of the sample and its ability to be generalized to a broader population.

7.1.4 Coding Scheme

The coding scheme for Study 4 was the same as Studies 2 and 3. Participants completed a total of 128 test items, 32 for each digit pairing.

7.2 Results

An ANOVA (Type II Wald $\chi 2$ tests) revealed statistically significant differences in the different training types ($\chi 2$ (219.52) = 3, $p < .05$), and the different digit pairings, ($\chi 2$ (114.23) = 3, $p < .05$). The Tukey Test revealed statistically significant differences when

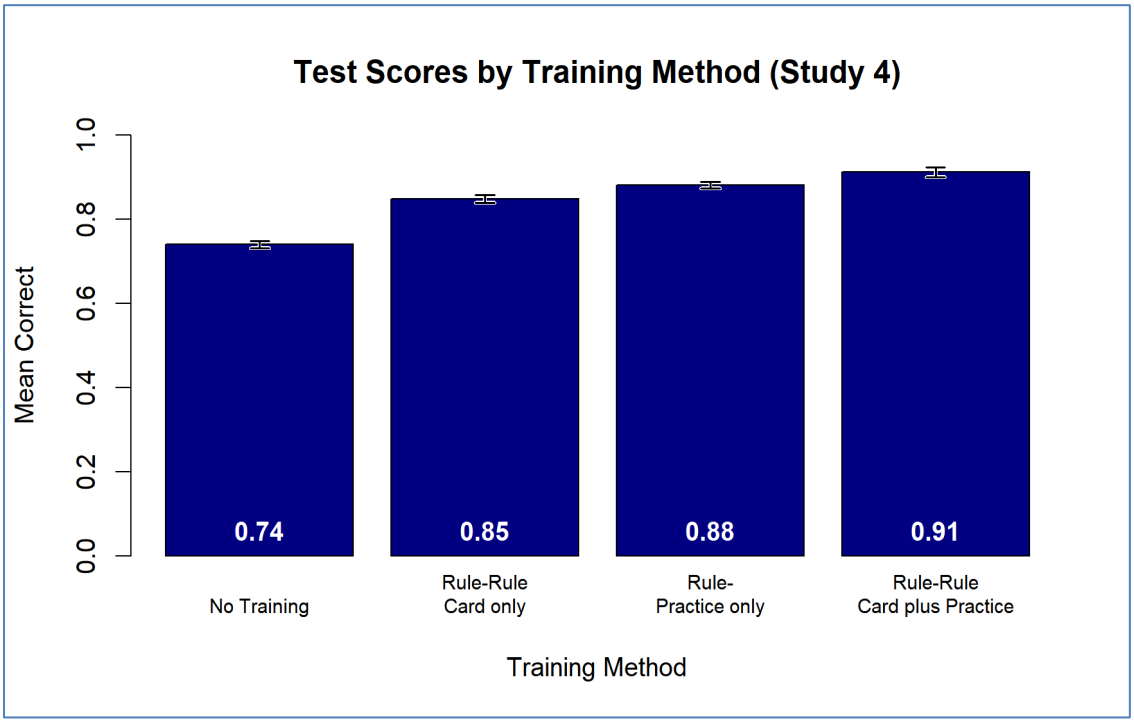
comparing all training types with each other, except for Explicit Rule Learning: Practice Only versus Explicit Rule Learning: Rule Card Only and Explicit Rule Learning : Rule Card and Practice.

R1: Are participants more accurate in predicting the output of the MNIST image classifier after receiving both components of Explicit Rule Learning (Rule Card + Practice with Feedback), Explicit Rule Learning (Rule Card only), or Explicit Rule Learning (Practice with Feedback only) training?

Participants performed similarly to Study 3. Removing ambiguous items and being more systematic and intentional by eliminating ambiguous items (digits where more than one rule can be applied), and by isolating the effects of Explicit Rule Learning (Rule Card + Practice with Feedback vs Rule Card Only vs Practice with Feedback Only) gave us a clearer picture of the effectiveness of the individual components of Explicit Rule Learning (Figure 7.1).

Figure 7.1

Study 4 Accuracy Results by Training Method. Error bars represent standard error.



7.3 Discussion

This study used No Training as a control condition, and also isolated the effects of Explicit Rule Learning, by testing the components separately and together. The results show that the best training was the full Explicit Rule Learning: Rule Card + Practice with Feedback. This led to participants to be better predictors of future cases with an average of 91% correct.

It should be noted that supporting the memorization of the Explicit Rule is helpful, as demonstrated by the 91% accuracy score where the participants received the full Explicit Rule Learning training (Rule Card + Practice with Feedback). The Practice with

Feedback was an effective tool to support the participants' ability to recall and apply the rules.

8 Discussion of MNIST Studies 1-4

The four studies evaluated the effectiveness of various methods and presentations of training (Table 8.1) for participants who were trained on an image classifier. The first study used an exemplar-based training method, which is often used to convey explanations of an AI/ML system's workings in XAI. Participants were shown exemplars that were factual (positive, i.e., correct classifications by the system) and counterfactual (negative, i.e., incorrect classifications by the system). The presentation of the training stimuli was either blocked or interleaved. The results showed that participants performed barely above chance when asked to predict the output of the classifier, indicating that this training was only a slightly effective way to train participants of these systems.

Table 8.1

Comparison of Conditions and Digit Pairings for Studies 1-4

Study	Conditions	Digit Pairings
1	• Positive Interleaved • Negative Interleaved • Positive + Negative Interleaved • Positive + Negative Blocked	Positive: **0** classified as **0** **6** classified as **6** **1** classified as **1** **5** classified as **5** **2** classified as **2** **3** classified as **3** **4** classified as **4** **9** classified as **9** Negative: **0** classified as **6** **6** classified as **0** **1** classified as **5** **5** classified as **1** **2** classified as **3** **3** classified as **2** **4** classified as **9** **9** classified as **4**
2	• No Training • Positive + Negative Interleaved • Explicit Rule Learning: Practice with Feedback only • Explicit Rule Learning: Rule Card + Practice with Feedback	Positive: **0** classified as **0** **1** classified as **1** **3** classified as **3** **4** classified as **4** Negative: **0** classified as **6** **1** classified as **5** **3** classified as **2** **4** classified as **9**
3	• No Training • Positive + Negative Interleaved • Explicit Rule Learning: Practice with Feedback only • Explicit Rule Learning: Rule Card + Practice with Feedback	
4	• No Training • Explicit Rule Learning: Rule Card only • Explicit Rule Learning: Practice with Feedback only • Explicit Rule Learning: Rule Card + Practice with Feedback	

Study 2 had a simplified collection of stimuli, adding a control condition (No training). Additionally, there was only one exemplar-based condition, and a rule-based training that was complemented with exemplars was introduced. The rule-based training proved to be a more effective method, whereby participants were better predictors of the system's classifications.

Studies 3 and 4 were similar to Study 2, but used more curated and intentionally selected stimuli, where each stimulus applied to one rule and one rule only. The explicit rule-based training was refined, as per Cognitive Tutorials for AI guidelines, and the structure for Rule Cards and Practice with Feedback was finalized.

Finally, in Study 4 the comprehensive Explicit Rule Learning: Rule Card + Practice with Feedback was compared to Explicit Rule Learning: Rule Card Only, Explicit Rule Learning: Practice with Feedback, and the control condition (no training). The accuracy of the participants' predictions of the system's classifications was upwardly mobile through the progression of the four studies and will be discussed next.

Test accuracy scores (percent correct) were calculated based on the correctness of participants, who were asked to predict the classifications made by the system (Table 8.2). A gradual improvement was made from Study 1 through Study 4. The exemplar-based training was not as effective as the rule-based training, which was global and explicit, and contained probabilistic, verbalizable rules complemented with factual and counterfactual exemplars. Being systematic about stimuli selection and curating the

stimuli selection (eliminating duplicates and ambiguous stimuli) led to higher test scores. The complete Explicit Rule Learning (Rule Card + Practice with Feedback) was the most effective training, as demonstrated by higher accuracy scores. Even isolating the components of the Explicit Rule Learning was more effective than exemplar-based training.

Table 8.2

Accuracy on test items for Studies 1-4, with a description of the studies

Study	**Negative Exemplars**	**Positive Exemplars**	**Neg./Pos Interleaved**	**Neg./Pos Blocked**	**No Training**	**Explicit Rule: Complete**	**Explicit Rule: Practice Only**	**Explicit Rule: Train Only**
1	0.56	0.51	0.56	0.55				
2 *1 direction (1-1, 1-5, eliminate 5-5, 5-1), Updated instructions*			0.69		0.65	0.73	0.73	
3 *Eliminate ambiguity, each test item only fits 1 rule, Clean up wording of instructions*			0.81		0.73	0.92	0.87	
4 *Separate Explicit Rule Learning components*					0.74	0.91	0.88	0.85

We observed the Explicit Rule Learning method as an effective approach to teach learners an AI/ML system in a laboratory environment. The next goal was to apply Explicit Rule Learning to a more sophisticated AI/ML domain, where the training and stimuli would be representative of a real-world, naturalistic environment in a complicated intelligent software system. We identified the Tesla Autonomous Vehicle (Full Self-Driving) domain for the next phase; this system uses a more advanced neural net AI.

9 Tesla FSD Study 5

The Tesla full self-driving system was the platform for the final study. This system was chosen to further explore the effectiveness of Explicit Rule Learning as a viable training method in a more sophisticated and complicated AI/ML system.

9.1 Tesla FSD

Tesla introduced full self-driving functionality (SAE Level 2; SAE International, 2021) to a limited number of drivers in October, 2021. In addition to its Autopilot features (traffic-aware cruise control, autosteer, Navigate on Autopilot, Auto Lane Change, Autopark, Summon, and Smart Summon) the FSD adds Traffic and Stop Sign Control and Autosteer capabilities (https://www.tesla.com/support/autopilot).

Tesla FSD uses a neural network, recursively trained with data received from Tesla FSD beta testers (https://www.enterpriseai.news/2023/03/08/how-tesla-uses-and-improves-its-ai-for-autonomous-driving/). Cameras on the vehicle perceive its current state and environment, and use this, along with data on which it has been trained, to achieve goals. The goals include planning and implementing a path towards a destination, with consideration to safety, time efficiency, and with attention to the dynamic presence and motion of objects in its environment (lane lines, curbs, traffic signs, other vehicles, pedestrians, etc.).

The ultimate goal of autonomous vehicles is to replace the human driver, with the vehicle having complete control of the traversal, making all the required decisions, and with the

goal of getting to a destination safely, legally, and efficiently. Although completely driverless vehicle functionality is being tested presently, and even with the great technological advancements that have been made thus far, currently, the human driver is still needed to perform a supervisory role. In this role, the human driver is tasked with remaining vigilant, ready to take over as needed.

The role of the human driver is to perceive, anticipate and respond to potentially adverse situations. If a human driver is meant to take over control of the autonomous vehicle "as needed", they must be well-informed of the cues and preceding actions made by the autonomous vehicle that indicate an impending adverse situation. In order to know when to take control over the autonomous vehicle, the human driver must have a deep understanding of the autonomous vehicle's AI system, where it succeeds and fails, and the boundary conditions that may change the system's response. Optimally, the human should also have an understanding of the goals of the system, as well as the logic and rationale used to achieve those goals. However, the compendium of these factors, including all positive aspects, negative aspects, system logic, rationale, and goals, is not readily available to drivers.

The entity with the most intimate knowledge of these factors might be the developer of the system. The proprietor has a team of designers, programmers, internal testers, etc. Naturally, technological advances, capabilities, and successes are publicly announced to garner support for the product. However, understandably, it is not in the best interest of the proprietor to advertise limitations, failures, and boundary conditions of the system.

Even if the system complies with all legally mandated regulations and standards, the act of driving from point A to point B is a complex task, with many critical junctions, challenging even the most capable of systems.

A complete understanding of the system, which is crucial for the human driver supervising the autonomous vehicle, includes all factors, whether positive or negative. Knowing the capabilities of a system is helpful. However, the limitations of a system might provide more useful information to new drivers of autonomous vehicles. Yet, limitations are not as easily discoverable in traditional support documentation and resources. The driver might start their discovery of the system's inner workings with the proprietor and their documentation. Additionally, however, they must also seek out other sources of information and support in order to gain a more complete understanding of the system.

One logical source for a complete evaluation of the system's capabilities and failures might be drivers who have used the system. In Tesla's 2023 Q1 Investor Relations report, Tesla boasts that 150,000 million miles have been driven by FSD beta testers (https://digitalassets.tesla.com/tesla-contents/image/upload/IR/TSLA-Q1-2023-Update). These drivers are not beholden to the proprietor, and offer impartial, unbiased reports on the system's performance. These pioneers are the first public cohort of beta testers, and their reports offer a rich collection of descriptions of the Tesla FSD system, its capabilities, failures, and conditions which may change the response of the autonomous vehicle. The typical autonomous vehicle driver generally has a large amount of driving

experience, is confident with their computer expertise, and is interested in how automation works (Dikmen and Burns, 2016). This unique perspective gives the drivers the ability to report pertinent details about events they've experienced with the FSD system, and also, a knowledge-based interpretation of the causes of the event, and the possible rationale and logic used by the system, as well as possible workarounds.

It seems natural that these drivers turn to a digital interface to communicate these experiences, such as the Internet and Tesla-related digital hubs to report their experiences, and to seek information on the experiences of other beta testers. One such hub is threaded social media (Linja et al., 2022).

Research done by Mamun (2023) demonstrated that Collaborative Explainable AI (CXAI), in the form of threaded social media, can be used as a non-algorithmic XAI, as it satisfies many of the goals algorithmic XAI developers seek to achieve. The abundant source of social media posts contains experiential reports from drivers in a naturalistic setting and provides rich information on events the drivers experienced while supervising the Tesla FSD system. This includes the aforementioned factors of the system's capabilities, failures, anticipatable cues that indicate an impending adverse action made by the autonomous vehicle, the danger level or relative consequences of problems, and descriptions of the drivers' mental models and their interpretation of possible reasons for the vehicles decisions and actions.

In a recent study, researchers demonstrated the feasibility of using threaded social media posts as a basis for CXAI (Linja et al., 2022). The study also resulted in a list of the most common failures and limitations of the Tesla FSD system reported by Tesla FSD beta testers as they interacted with the autonomous vehicle. These posts were examined for their content, but also evaluated for reports of events that matched (or didn't match) the expectations and mental models of the drivers. This list of social media posts was coded for themes and rate of occurrence.

Table 9.1 contains a list of the final themes, as coded by two independent coders (Cohen's k = 0.92, indicating a high of agreement; Cohen, 1960). This information, from actual Tesla FSD beta testers, provides a well-rounded corpus for consideration when developing training content. This includes all factors (positive and negative) that, when taken as a whole, can be analyzed and rated for levels of importance, danger, and can be indicative of the system's underlying framework. The result is a refined list of operational principles, cues to be aware of, unexpected maneuvers made by the automated vehicle and patterns which have been identified. This final product is the basis for the training content for Study 5.

Table 9.1

List of coded categories and problems from social media reported by Tesla FSD drivers

Label	**Description**	**Example(s)**	**Count: Both (Either) Coder**
Lane usage	Unexpected lane usage or lane maintenance	Hug center of road/go straight from turn lane	71 (78)
Stopping	Unexpected stopping or slowing down	Phantom braking/stop half a block before the stop sign	43 (45)
Jerky ride	Unnecessary/sudden starts/stops	Jerky turns/brake or accelerate with a sudden jerk	22 (22)
Timidness	Timid Approach	Timid to commit to turn/turn-taking at 4-way stop	20 (26)
Impeding	Impeding other vehicles	Almost impacting another vehicle/following too close	17 (23)
Obstacle speed	Approach impending obstacle too fast or accelerating too fast	Excessive speed at a turn/roundabout	12 (16)
Turning	Improper turning	Wide turns, tight turns, blocking vehicles when turning	12 (13)
Steady speed	Driving too fast/slow for conditions	Unexpectedly driving too fast/slow steadily	12 (12)
Signaling	Improper turn signal usage	Failure to apply, phantom application, wrong turn signal, applies late	11 (12)

Pathfinding	Mismatch between tentacle and actual path	Did not follow GPS route as displayed on screen	8 (10)
Warnings	Inappropriate false system warnings	Inappropriate/false forward collision warnings	8 (9)
Disengagement	Vehicle initiated disengagement or stopped working and did not proceed	FSD stops working/vehicle stops without apparent intent to proceed	6 (7)
Mapping	Unaware of current map configuration	Obsolete/incorrect map data	4 (6)
Camera	Unexpected screen, visualization, camera rendering, or interpretation	Misjudging position of other vehicles/objects	3 (5)
Recognition	Inability to recognize non-road entities	Parking lots, driveways, residential area entrances	3 (4)
U-turns	Problems making U-turns	Avoid/disengage; U-turn turned into a left turn	2 (2)

9.2 Goals and Research Questions

Explicit Rule Learning was proven to be an effective training method for an ML image classifier, guiding the participants through the system's rationale, strengths and weaknesses, capabilities and limitations, and boundary conditions that changed the output. However, it was unknown if Explicit Rule Learning was robust and translatable to a more complicated AI system in a real-world application. Additionally, having established the formula for developing an Explicit Rule Learning tutorial, an investigation was needed to determine whether or not a tutorial could be developed for a

more complicated AI system without fundamentally changing the Explicit Rule Learning method.

The Tesla FSD domain provides a rich setting to test Explicit Rule Learning. The underlying system is sophisticated and intelligent, the environment is dynamic and naturalistic, learners can relate to the need to understand the system's framework and rationale prior to using the system, and the experiences of drivers testing the FSD system is plentiful in Tesla FSD-specific social media platforms.

The goal of the final study was to create Explicit Rule Learning training on this sophisticated intelligent system and test its viability. The research questions were as follows.

R1: Can Explicit Rule Learning be adapted to a more sophisticated intelligent system such as the AI system used by full self-driving autonomous vehicles? Can verbalizable rules be identified that would accelerate the proficiency of participants of such an advanced intelligent system? Knowing that it is not possible to obtain a true probability, is it possible to identify rules that are most likely to occur and make a difference in a participant's performance when predicting the output of the system?

R2: Are participants more accurate in predicting the output of the Tesla FSD system with Explicit Rule Learning versus the control condition (no training)?

To address these questions, we conducted a controlled experiment using a within-subjects experimental design. We developed four rules that were used to train participants in the intelligent software system. The participants were trained on two rules each and tested on all four rules in this within-subject study.

9.2.1 Rule Content

The stimuli for Study 5 came from an existing corpus of statements (Linja et al., 2022). The statements were social media posts made by Tesla drivers who used the full self-driving feature. The drivers posted about unexpected responses or actions made the autonomous vehicle, problems, safety issues, illegal maneuvers, negative experiences with the decisions made by the autonomous vehicle, and possible explanations and speculations about the FSD system's rationale. The posts were from the first six weeks of the widespread FSD beta release beginning in October, 2021.

The corpus of posts was parsed into 273 statements (each containing one event) and coded (Cohen's $\kappa = 0.92$; Cohen, 1960) to determine themes and the frequency of their occurrence. This resulted in 19 themes, along with their base-rate (Table 9.1). These thematically coded results were reviewed, and 4 themes were selected as candidates for the content of the Explicit Rule Learning training. The selection process was based on two factors. First, the candidates for Explicit Rule Learning were based on the ability of

the issue to not only instruct the learner of one FSD feature, but also the ability to generalize the knowledge to the global rationale and logic of the FSD system. Secondly, issues that were considered most dangerous, or had a higher rate of occurrence were given priority.

9.2.2 Rule Cards

Once the issues were identified, the rules were developed, mirroring the circumstances as reported by the Tesla FSD drivers and the outcome of the situations. The four rules were used to create Rule Cards (Appendix A). As in the previous studies, the Rule Cards contained the textual rule descriptions and summaries, textual and visual depictions of the base rate of occurrence, and factual and counterfactual exemplars in the form of static images and videos.

Following is a description of a commonly identified event that occurred with Tesla FSD drivers, the creation of an explicit rule, and the transformation of the rule into Explicit Rule Learning training material.

One of the more frequently reported issues was situations where the Tesla FSD was driving on a 2-lane residential road, without lane lines. The Tesla FSD tended to “hug” the center of the road, driving in the middle, straddling the oncoming lane and the right lane.

1. Rule development
    - Description: “In cases where the Tesla FSD is driving in a residential area without lane lines, it will most likely drive in the center of the road, straddling the oncoming lane and the right lane.”
    - Possible Rationale: “Without lane lines painted, the Tesla may erroneously consider the edges of the road (i.e., curbs) as its lane (right and left) boundaries, therefore centering itself on the road rather than in its own lane on the right side of the street.”
    - Possible Outcomes: “When the Tesla FSD is driving in the center, if an oncoming vehicle is approaching, the driver supervising the Tesla will need to take control of the vehicle and shift right, or the oncoming vehicle will be forced to shift out of the Tesla FSD’s trajectory.”
    - Base Rate: It would be theoretically possible to identify the true base rate of occurrence. For example, given access to a Tesla autonomous vehicle with FSD functionality enabled, one could drive on a sufficient number of residential roads *with* and *without* lane lines, and identify the number of times the Tesla FSD drove in the middle, straddling both the oncoming lane and the right lane, and the number of times it drove in the right lane as a human driver would expect it to. However, there would be insuperable obstacles preventing this from being accomplished. For example, first, one would need access to a Tesla FSD vehicle. Next, one would need to identify quite a few relevant roads (i.e., residential roads *with* lane lines painted, and residential road *without* lane lines painted) to identify a true probability. Additionally, the relevant roads and probability would have to

be generalizable to other residential roads, in other towns in other states, both densely populated and not, and with a considerable number of other variations that occur on roads.

Due to these insurmountable restrictions, and having performed due diligence by reviewing the data, it was decided that for this study, probabilities would be determined with some degree of analysis and defined as “low” (least likely to occur, less dangerous), “medium”, and “high” (most likely to occur, most dangerous). These were represented in a manner consistent with the driving domain – a “green”, “yellow” or “red” light.

This was an interesting factor we identified when testing the adaptability of Explicit Rule Learning in a more complicated AI system.

2. Rule Card Development
    - Once the rule content was developed, the next step was to collect static images and/or videos to depict factual and counterfactual instances of the rule. This was done by searching the Internet first. In most cases, we were able to find the necessary rule representations online, remove any identifiable information, and use these as stimuli for the study. In some cases, where we could not find rule representations, or the minimum quantity required to represent the rule, the images and/or videos were created manually using computer-generated visualizations. Here, the priority was to collect images/videos that represented the

rules the best; the format of the file (image, video, computer-generated graphic) was of less importance.

3. Rule Card Presentation
    - In this study, all of the tasks were online. Therefore, a digital file (.png) was created. Figure 9.1 is a sample Rule Card for this example.

Figure 9.1

Sample Rule Card for Study 5

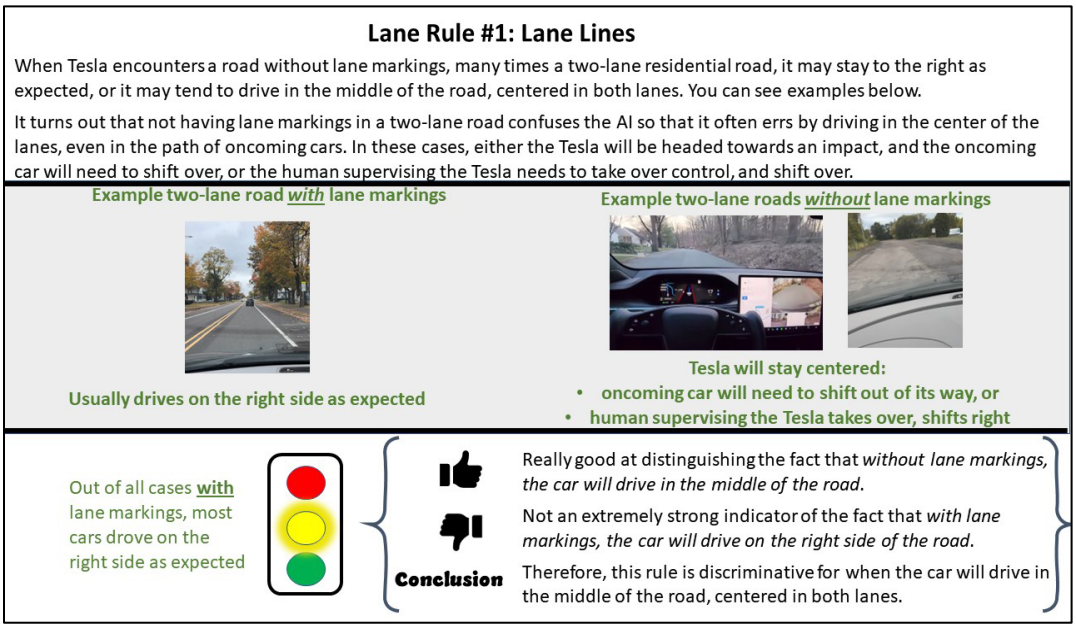


Table 9.2 and Figure 9.2 illustrate the standard process for the identification of a principle, and the subsequent rule card developed for the MNIST studies and the final Tesla FSD study.

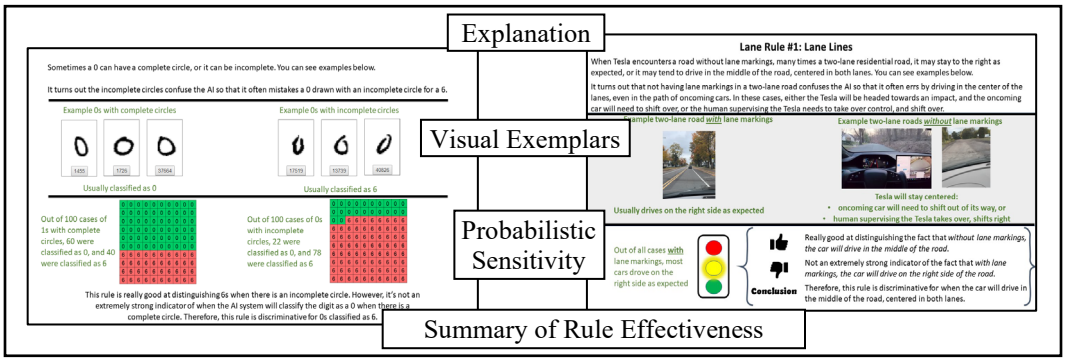
Table 9.2

Comparison of Rule Development for MNIST Studies 1-4 vs. Tesla FSD Study 5

	MNIST	**Tesla FSD**
Rule Development	When a 0 (zero) is drawn without closing the circle, the classifier often misclassifies it as a 6.	On a residential road without painted lanes, the Tesla FSD often drives in the middle of the road, straddling both the oncoming and right lanes.
Possible Rationale	The incomplete circle is more similar to the 6s which the classifier has been trained on than the 0s.	Without lane lines painted, the Tesla may erroneously consider the edges of the road (i.e., curbs) as its lane (right and left) boundaries, therefore centering itself on the road rather than in its own lane on the right side of the street.
Possible Outcomes	The 0 will be misclassified as a 6.	When the Tesla FSD is driving in the center, if an oncoming vehicle is approaching, the driver supervising the Tesla will need to take control of the vehicle and shift right, or the oncoming vehicle will be forced to shift out of the Tesla FSD's trajectory.
Base Rate	Out of 100 0s that were drawn with a closed circle, 60 were classified as 0. Out of 100 zeros that were drawn without closing the circle, 78 were misclassified as 6.	An image of a yellow traffic light indicated a medium base rate for this occurrence.

Figure 9.2

Comparison of Rule Cards for MNIST Studies 1-4 vs. Tesla FSD Study 5



9.2.3 Practice with Feedback and Test Stimuli

Next, the stimuli for the Practice with Feedback and Test items were created. All of these were created from videos found online, mostly from https://www.youtube.com/. There are many Tesla FSD drivers that recorded themselves testing the FSD beta version, with cameras focused on the inside of the vehicle, showing the dashboard complete with a display monitor visualizing the navigation system and video captured by the FSD's cameras. Additionally, these recordings contain adjacent frames that also display the driver's point of view, showing the road in front of them and a limited peripheral view. As the drivers traverse the various roads (rural, urban, some sparsely marked, some complicated road configurations), the video displays both the inside of the vehicle and the surrounding environment from the driver's perspective.

For this study, these YouTube videos were poured through and inventoried. From this list, clips were selected that represented factual and counterfactual instances of each of the rules. These clips were processed individually using video editing software.

First, any personal information, or information that could lead to the identification of the driver and/or poster of the video was blocked or blurred out. Next, the videos were enhanced with text and shapes. Most of the videos represented the rules completely; however, some videos were ambiguous, and the participants needed to be guided. For example, perhaps there were cues that needed to be pointed out (e.g., "There's a car coming from the left"), or perhaps some areas of the screen were more relevant than others ("Notice the navigation route indicates an upcoming left turn"). In these cases, text was overlayed on the video to help the participants, and the relevant areas were made more salient with boxes or circles outlining them.

The final videos were approximately 15-25 seconds long. At some point in the video, there was a "freeze frame", and the participant was asked "What will happen next?". For example, a video for the rule described above showed the Tesla FSD vehicle driving in a residential area with no lane lines painted. The Tesla FSD is driving in the center of the road, straddling the oncoming lane and the right lane. Suddenly, an oncoming vehicle turns onto the road, coming toward the Tesla. At this point in the video, it is clear that the Tesla is impinging on the oncoming vehicle's lane, and if both vehicles continue without modification, they will collide. The video freezes, and a text overlay is shown: "What will happen next?"

In the Practice with Feedback portion, the participant responds to multiple choice or yes/no questions. In the test items, the participant responds by typing their answer in a short-answer field.

9.3 Questionnaires

The proficiency of the participants in these studies was rated by analyzing test results. However, it is possible that other factors influence the learning process. For example, Eccles and Wigfield (2020) posited that participants' perceived expectancy, value, and cost of the material and learning process influences the learning outcome. In order to have analysis on factors that may influence learning, and to learn more about factors that might provide patterns and nuances that are not visible by quantitative results alone, participants were asked to respond to several questionnaires. These questionnaires are not the focus of the study, rather, the results were used as post hoc investigatory tools to support the main focus of the study, Explicit Rule Learning. They were also used to inform future studies.

A description of the questionnaires follows. In brief, the following questionnaires were given to participants: Demographics, Cellphone Usage While Driving, Trust in Automation, and User Experience. Additionally, a subset of participants participated in an oral interview, where they were asked questions about the relevancy of the training as a consumer of the Tesla FSD system.

9.3.1 Demographics/ Cellphone Usage While Driving

At the onset of the study, participants were asked to provide demographic information to determine the representativeness of the participants for generalization purposes. This included their gender, current age, the age at which they obtained their driver's license, and the estimated number of miles driven in the past 12 months. The demographic responses were summarized.

Participants were also asked to provide two ways they use their cellphones while driving. Research has shown that cellphone usage while actively driving can distract drivers, leading to adverse events (Atwood et al., 2018). These effects are not lessened when some of the cellphone functionality is taken over by automation. For example, although an autonomously driven vehicle takes over the responsibility for some of the cellphone functionality (e.g., GPS/Navigation), other cellphone functions (e.g., answering a phone call) still have a negative impact, distracting drivers of automated vehicles. Although the human driver is in a supervisory role of an automated vehicle, there are still attention requirements, as they still need to monitor the current and potential states, and be ready to take over as needed. Cellphone usage while operating an automated vehicle increases the time to take over (Merlhiot and Bueno, 2022; Neubauer et al., 2012; Zhao et al., 2022).

In summary, if drivers are using their cellphone solely for tasks such as GPS or navigation, this type of distraction would be eliminated, as the autonomous vehicle takes over this responsibility. If drivers are using their cellphone for other things, such as texting, or skipping over songs in a music app, this distraction would still exist while

operating an autonomous vehicle. In order to determine the most frequently occurring purpose of using a cellphone while driving, our first research questions was:

R3a: What are the most frequently reported tasks performed on a cellphone while driving? Are they functions that would be subsumed by the autonomous vehicle (such as navigation), or apps that would not be taken over by the autonomous vehicle (such as texting or skipping songs), and therefore still potentially distract the driver?

9.3.2 Trust in Automation Questionnaire

Previous researchers have suggested that trust plays a critical role in the human-automation interaction and may be a determining factor on the human's reliance upon the automated system (Hoffman et al., 2018; Wang et al., 2021). Lee and See (2004) stated that with trust, a human will rely on automation. However, when lacking trust, humans will reject the automation. However, a human's trust in an automated system is not binary, either present or absent. Nor is it static and unfluctuating. Trust modulates with certain events (Yu et al, 2017), at times increasing, and at other times decreasing. For example, repeated failures by the automated system decrease the trust level of the human, but this decrease can be recovered by showing successes of the automated system (Sauer et al., 2016).

Reliance on human-automation teaming varies depending on the learners' dispositional, situational, and learned trust (Hoff and Bashir, 2015), which can subsequently positively (or negatively) affect the learners' motivation to engage with and learn about the system.

An intelligent software system's explanation for a specific output, and the content and presentation of a system's training material can have an effect on a learner's interaction, with some more trust-building than others (Wang et al., 2016).

In order to measure the dynamically changing trust level in this study, the Körber Trust in Automation Questionnaire (Körber, 2018) was given to participants at three points during the study (at the beginning of the study, after the training but before the test, and post-test). This questionnaire consists of 19 self-reported trust statements, to which participants respond with their level of agreement or disagreement using a 5-point Likert scale. The questions gauge the participants' level of trust with respect to Reliability/Competence, Understanding/Predictability, Familiarity, Intention of Developers, and the Propensity to Trust and Actual Trust of the automated system.

The responses to this questionnaire were analyzed at each of the three points in time and compared temporally. The research question for the Trust in Automation Questionnaire is:

R3b: Will the participants' trust in automation decrease after being shown the failures of the Tesla FSD system despite their training benefit? Will there be a correlation between the Explicit Rule Learning test scores and the trust ratings?

9.3.3 User Experience Questionnaire

A User Experience Questionnaire was presented to the participants post-test as well, where the participants responded to questions (via a 5-point Likert scale) about their

perception of the reliability, validity, and source of the training material. This questionnaire was adapted from a parallel study conducted by TI Mamun, who is also in the Veinott/Mueller Lab. The results of the questionnaire will be summarized. The research questions for the User Experience Questionnaire are:

R3c1: How will Tesla FSD novices rate the effectiveness of the training?

R3c2: Will the novices believe the training came from an authority in the domain (i.e., Tesla) or from peer end users (i.e., other Tesla FSD drivers)?

9.3.4 Consumer Application: Oral Interview Questions

A subset of 13 participants were asked to participate in an oral interview after completing the online portion of the study (Appendix E). The additional questions added value to our results by providing mixed methods quantitative and qualitative data, and also demonstrated a broader application of Explicit Rule Learning in a real-world context (i.e., a consumer considering the purchase of an autonomous vehicle, their evaluation of the training and whether or not it helped them to understand the system better, whether or not it affected their trust and knowledge in the autonomous vehicle domain, and how it might affect their decision to purchase an autonomous vehicle.) Specifically, participants were asked to evaluate the training from the perspective of a potential Tesla FSD consumer, bridging the in-laboratory training to the real-world, reflecting on the effectiveness of the training as an actual "in the wild" training tool. Our research questions for this portion were:

R4a: What resources might a new Tesla FSD driver go to in order to find training material on the system?

R4b: With regards to the specific functionality the participants were trained on in the study, was the training sufficient in that that they would feel comfortable using that specific functionality of Tesla FSD? In what way(s)?

R4c: Does the training affect a potential Tesla FSD consumer's likelihood of purchasing the system? Does the awareness of failures identified in the training deter purchasers?

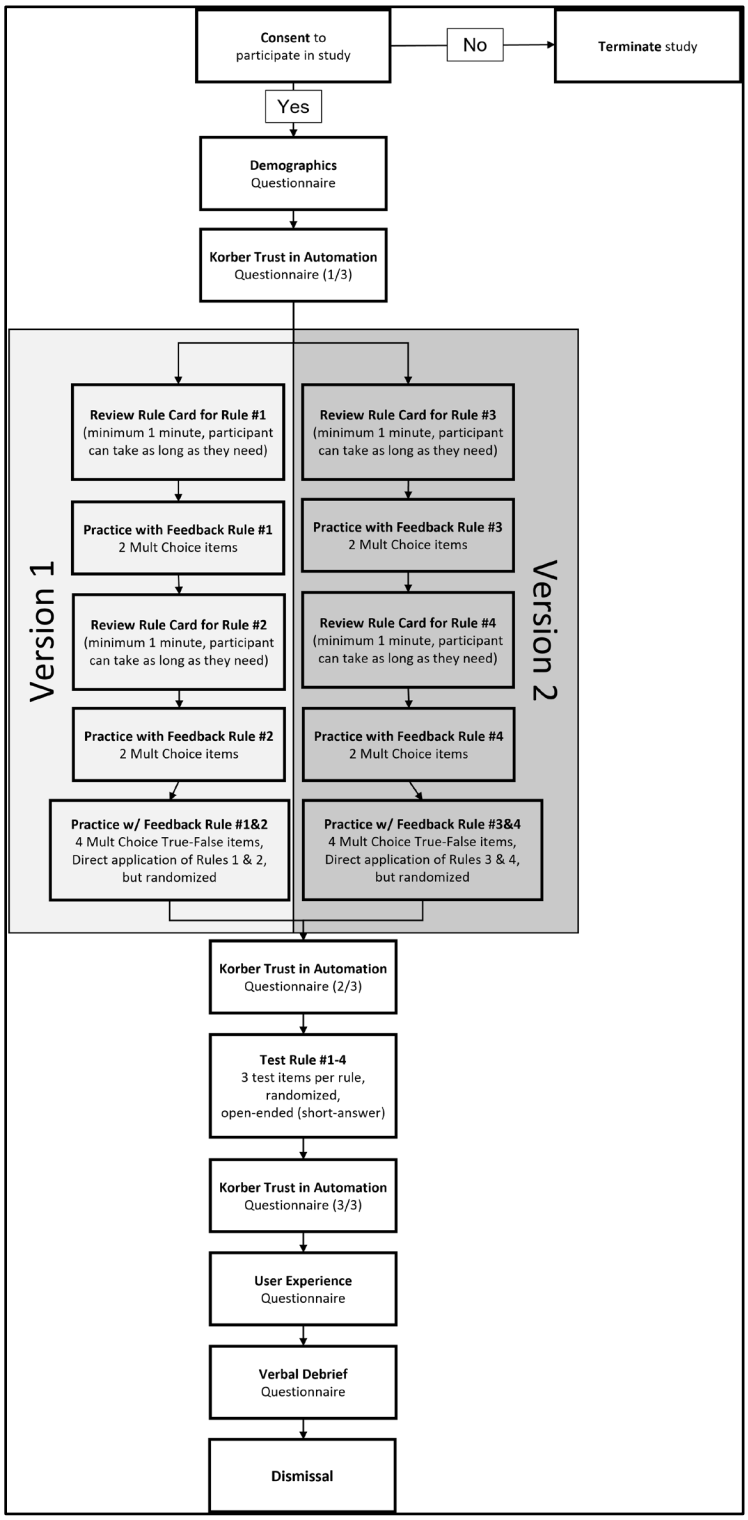
R4d: In what ways do participants understand autonomously driven vehicles now, after the training, that they did not previous to the training?

9.4 Study Flow Summary

The final survey flow for the study is shown in Figure 9.3. The participants were alternately assigned to a version in which they received Explicit Rule Learning for two of the four rules and were tested on all four rules.

Figure 9.3

Tesla Survey Flow Illustrating Participant Activities in Study 5



9.5 Coding Scheme

There was a total of 12 test items, 3 for each rule. Results were coded for accuracy (0 points for “incorrect” and 1 point for “correct” responses). The short-answer open ended responses were qualitatively coded by two independent coders. The Trust in Automation and User Experience Questionnaire responses were coded quantitatively. The remaining questionnaires (demographics, cellphone usage, and Consumer Application Oral Interview Questionnaires) were coded by at least two independent coders.

9.6 Analysis

The experimental and control conditions, and the four different rules were tested for statistically significant differences. Additionally, accuracy scores were compared between the conditions and rules.

9.7 Method

9.7.1 Participants

9.7.1.1 Power Analysis

The power analysis came from two sources. First, we ran an informal pilot using a portion of the Tesla FSD stimuli, with an abbreviated form of Explicit Rule Learning; a power analysis was run with these results. Additionally, a parallel study was run by a researcher in the same Cognitive and Learning Sciences lab, which had similar dependent measures. Assuming the pilot results represent the truth, and taking into consideration the results from the parallel study, it was determined that at least twenty-five participants were needed to find and effect for the within-subject study. Additionally, the power

analysis determined that three test questions were required per rule to obtain the statistical effects desired for the study.

9.7.1.2 Participants

Forty-seven college-aged students (62% male, 36% female, 2% non-binary/non-conforming), average age of 20 years (SD = 1.42) completed the within-subject 45-minute study in exchange for Introduction to Psychology course credit. The study was completed online via the Qualtrics survey platform. All of the participants were licensed drivers and drove between 5 and 30,000 miles in the past 12 months (M=6,635 miles, SD=7,729, Median=4,000).

9.7.2 Tesla FSD Prediction Task

After completing the Demographic, Cellphone Usage While Driving, and Körber Trust in Automation Questionnaires, participants advanced to the training portion of the study. In an alternating order, half of the participants were trained on Rules 1 and 2, and half on Rules 3 and 4.

Using a balanced Graeco-Latin square (Table 9.3) with 4 conditions each (2 rules: control, no training and 2 rules: experimental, Explicit Rule Learning training), each of the participants were alternately assigned to one of two versions of the study.

Table 9.3

Study 5 Graeco-Latin Square with Rules and Training Conditions

Version	Condition/ Rule 1	Condition/ Rule 2	Condition/ Rule 3	Condition/ Rule 4
1	Explicit Rule Learning (Lane_1)	Explicit Rule Learning (Lane_2)	No Training (Approach_3)	No Training (Approach_4)
2	Explicit Rule Learning (Approach_3)	Explicit Rule Learning (Approach_4)	No Training (Lane_1)	No Training (Lane_2)

Participants completed the study at the location of their choice on www.Qualtrics.com. They were given the following instructions:

> "Imagine you are sitting behind the wheel of an autonomous vehicle. The full self-driving AI system is in control, and as the human driver supervising the system, you'd be able to take control of the vehicle at any time you think it's necessary. For example, you might feel that the autonomous vehicle is about to make a dangerous or illegal maneuver, so you'd be able to disengage the system and take over the driving at any moment.
>
> In this study, you will be shown videos from the perspective of the human driver supervising the autonomous vehicle, a Tesla. At some point the video will freeze, and you'll be asked to predict what will happen next."

Participants were then given a sample video with the following information:

> “Example: Play the video, and at the freeze frame, try to think about what the autonomous vehicle might do next.
>
> After each video, you'll be asked to respond with your answer about what you think will happen next. Let's begin!”

Next, the participant began the study, commencing with the Rule Card for the first rule. After the Rule Card, they were given two Practice with Feedback questions applying the first rule. They responded to the multiple choice questions, and were given feedback either reinforcing their correct response, or explaining the proper application of the rule if they responded incorrectly. The responses for the questions were either 4-item multiple choice, where one response was correct, and three others were incorrect, or 2-item yes/no choices.

After the two practice items, they were presented with the Rule Card for the second rule, followed by two Practice with Feedback items. Finally, they were given four more Practice with Feedback items, and were told that either of the two rules they learned might be applied to these items. It was up to them to determine which rule applied, and the proper application of the rule to the scenario.

This concluded the training portion of the study, and the Körber Trust in Automation Questionnaire was given to them for the second time.

Next, the participants were tested on all four of the rules. They were given 12 test items, three videos for each of the four rules. The order of the videos was randomized for each participant, and had the same format as the training, where at a certain point there was a freeze-frame, and the participant was asked to predict what would happen next. For the test items, the responses were short-answer. Participants were given the following instructions:

> "Now you'll be tested on the material. After each video, please respond with your best guess of what will happen next. This will be short answer, not multiple choice or yes/no, and you're being asked to write a few words about what you think will happen next, not what you as the driver in control of the vehicle would do."

9.8 Results

An ANOVA (Type II Wald $\chi 2$ tests) revealed statistically significant differences in the experimental (Explicit Rule Learning) vs control (No training) conditions ($\chi 2$ (52.63) = 1, $p < .05$), and the different rules, ($\chi 2$ (14.59) = 3, $p < .05$). The Tukey Test revealed statistically significant differences when comparing the experimental and control conditions.

9.8.1 Prediction Accuracy

Each of the forty-seven participants completed 12 test items (three for each rule), for a total of 564 test items. One participant reported that they were unable to view one of the

12 test videos (presumably some technical difficulty), so that record was eliminated from the dataset, leaving a total of 563 test items.

To score the responses for accuracy, an Answer Key was developed, listing possible "correct" responses, which were given a score of one point, and possible "incorrect" responses, which were given a score of zero. Two independent raters scored a subset of test items in terms of correct (1 point) or incorrect (0 points). After several rounds of training and discussion, each rater scored each of the 563 test items for accuracy, achieving a high interrater reliability with a Cohen's Kappa = .81 (92% match on scores). The remaining 45 records, which the raters independently scored with opposing results, were discussed individually. A consensus was reached on all disagreements, resulting in total agreement for each of the 563 scores.

R1: *Can Explicit Rule Learning be adapted to a more sophisticated intelligent system such as the AI system used by full self-driving autonomous vehicles? Can verbalizable rules be identified that would accelerate the proficiency of participants of such an advanced intelligent system? Knowing that it is not possible to obtain a true probability, is it possible to identify rules that are most likely to occur and make a difference in a participant's performance when predicting the output of the system?*

Explicit Rule Learning was successfully adapted to the Tesla FSD domain. The content came from actual Tesla FSD drivers, and reflected actual events as reported by drivers in the real world. Using this content, we developed rules that contained if…then statements, factual and counterfactual exemplars, and the general probabilities associated with the

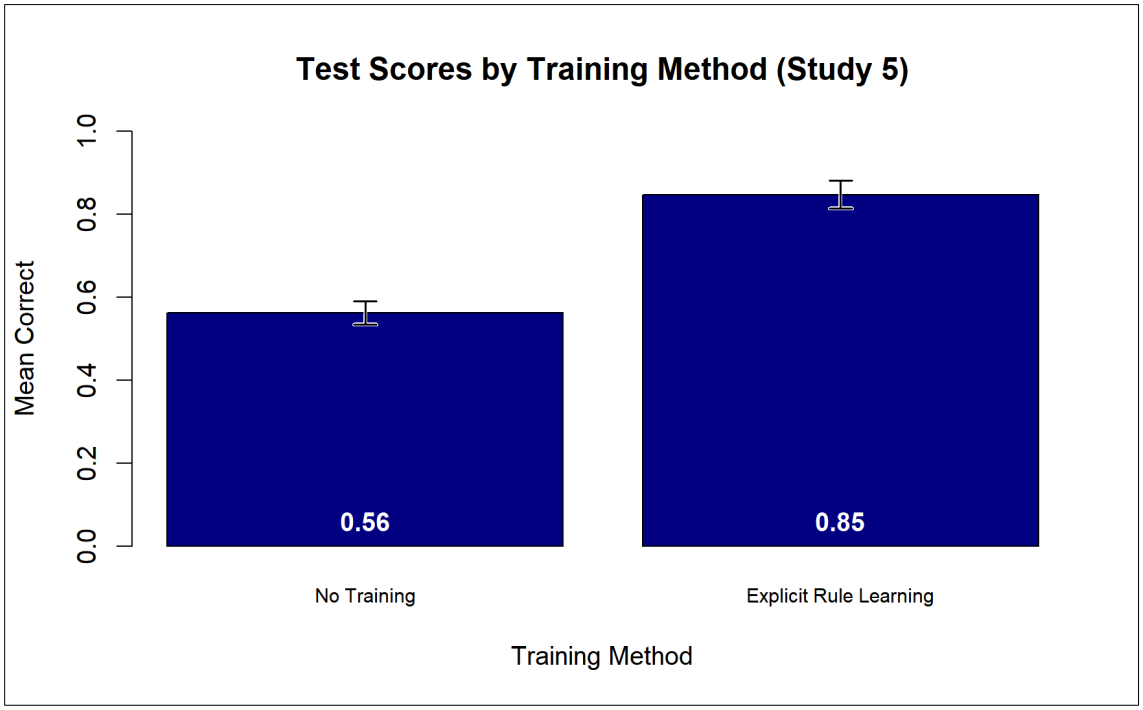
rule. Even without true base rates, it was possible to relate the impact severity of the rules to the participants.

R2: *Are participants more accurate in predicting the output of the Tesla FSD system with Explicit Rule Learning versus the control condition (no training)?*

Participants in the Explicit Rule Learning condition performed better than those in the No training condition (Figure 9.4).

Figure 9.4

Study 5 Accuracy Results by Training Method. Error bars represent standard error.



9.8.2 Questionnaires

9.8.2.1 Cellphone Usage

As described previously, cellphone usage can distract drivers, whether operating a manually driven, or an autonomously driven vehicle. The distraction is greater for those cellphone functions that are not subsumed by the automated system (i.e., GPS and navigation versus texting or skipping a song).

R3a: What are the most frequently reported tasks performed on a cellphone while driving? Are they functions that would be subsumed by the autonomous vehicle (such as navigation), or apps that would not be taken over by the autonomous vehicle (such as texting or skipping songs), and therefore still potentially distract the driver?

Drivers reported that they most frequently used their cellphones for functionality that is not taken over by an autonomously driven vehicle. The most frequently reported usage of a cellphone was listening to and controlling music or podcasts while driving. This task was mentioned by 39 (83%) of the 47 respondents. The second most frequently reported usage was maps and navigation, followed by phone calls. The complete list can be found in Table A.1, which is in Appendix A.

9.8.2.2 Trust in Automation Questionnaire

R3b: Will the participants' trust in automation decrease after being shown the failures of the Tesla FSD system despite their training benefit? Will there be a correlation between the Explicit Rule Learning test scores and the trust ratings?

Trust was measured at three points in the study: at the beginning of the study, after the training, and after the test. The overall trust decreased throughout the study (Table 9.4). This was expected for several reasons.

Table 9.4

Mean Trust Scores Pre-Study, Post-Training, and Post-Test

Time Point	Overall Mean (low=1, high=6)	Mean Trust for Above Average Test Scores	Mean Trust for Below Average Test Scores
Pre-study	3.07 (SD=1.02)	3.03 (SD=.96)	3.11 (SD=1.08)
Post-training	2.87 (SD=1.10)	2.99 (SD=1.10)	2.75 (SD=1.10)
Post-test	2.7 (SD=1.02)	2.73 (SD=1.08)	2.68 (SD=0.96)

First, the training content included descriptions of the system's failures (i.e., when the Tesla FSD fails to drive in the expected lane, and when the Tesla fails to proceed as expected after stopping at a stop sign or turn). As described in Section 9.3.2, showing participants a system's failures will lead to a decrease in trust. Secondly, each participant was trained in two rules, but tested on all four rules. After training, participants who were confident in their abilities to predict the system's output were challenged with test items upon which they were not trained. Confirmed by interview responses, this caused the participants to be surprised, as they realized they had gaps in their knowledge, which led to decreased trust.

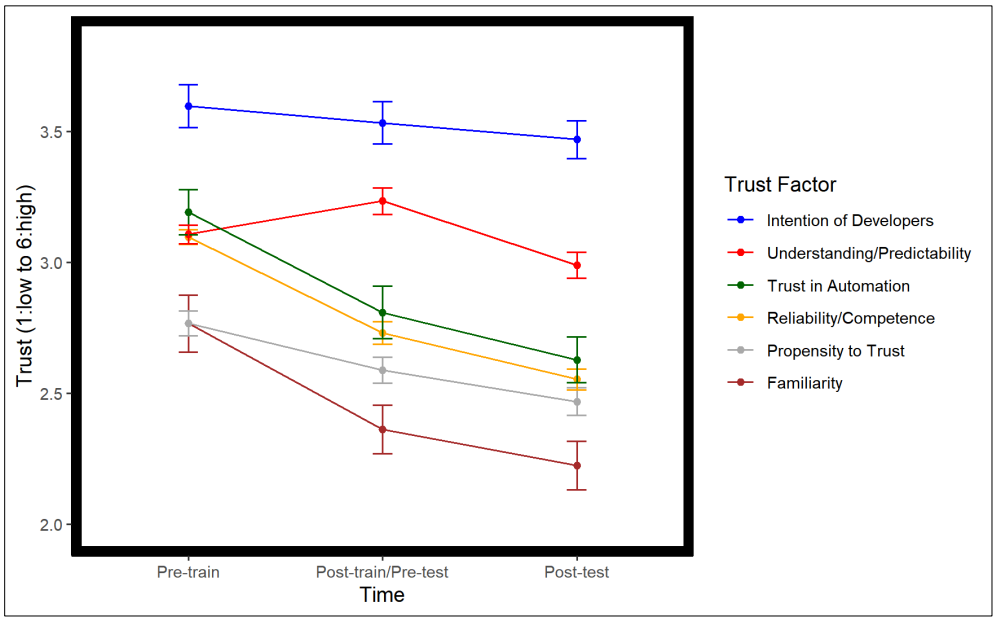
The trust ratings were also compared against accuracy scores. Although not by a significant amount, those participants achieving test scores above the average test score had the least amount of trust decrease from the Pre-study rating to the Post-test rating. While the average trust score decreased by 12% across all participants, it decreased 14% for those participants whose test scores were below the average score, and only 10% for those participants whose test scores were above the average score.

An ANOVA (Type II Wald $\chi 2$ test) revealed statistically significant differences in the overall trust between the three time points ($\chi 2$ (66.4) = 2, $p < .05$). There were no statistically significant differences between the three trust scores and the mean test scores of the participants ($\chi 2$ (2.5) = 1, $p = .11$), or the study version (Explicit Rule Learning training in two *lane* rules vs training in two *timid approach* rules) ($\chi 2$ (1.6) = 1, $p = .21$).

There were differences among the individual factors of trust (Figure 9.5); however, only some of the factors had statistically significant differences between the time points (for detailed results see Table C.1 in Appendix C). Interestingly, Intention of Developers and Propensity to Trust were the two factors that did not have statistically significant differences. The trust in the Intention of the Developers had the highest overall score, and trust in the Familiarity of the FSD system had the lowest overall score.

Figure 9.5

Study 5 Trust Scores Taken at Three Points. Error bars represent standard error.



A paper published by Endsley (2017) reported that although there was some minimal explanation on the Tesla FSD system at the purchase point, most of the learning was done autodidactically, or by using some other ad hoc method of gathering driver experiences (talking to other Tesla drivers, Tesla dealership informal training).

In a study with current, experienced Tesla FSD drivers (Mamun, 2023), when asked where they would go for training material, most drivers stated they would go to other drivers (via social media, network of acquaintances, etc.) for training and information. The drivers did not report that substantial training came from Tesla, the developer and authority of the FSD system.

However, the novices in the FSD domain, our participants, had a higher level of trust for, and a higher level of dependency on, the intention of the developers of an AI system. Additionally, their levels of Propensity to Trust automation remained at the same level. Körber (2019) defines this factor as the user's trust in the ability, benevolence and integrity of the automation developer. The novice drivers (i.e., the participants in our study) were more inclined to trust the developer and consult them as an authority in learning about the system. This will be discussed further in the results of the Consumer Application Interview discussion.

9.8.2.3 User Experience Questionnaire

R3c1: How will Tesla FSD novices rate the effectiveness of the training?

The results of the User Experience Questionnaire (adopted from Mamun, 2023) reveal that the participants rated the training favorably, leading to participants' confidence in their ability to predict the FSD's actions (Table 9.5).

Table 9.5

Mean levels of agreement for User Experience Responses

Statement	Level of Agreement (disagree=1, agree=5)	Level of Agreement for Above Average Test Scores	Level of Agreement for Below Average Test Scores
I was able to identify the autonomous vehicle's subsequent actions for each video.	3.3 (SD=.8)	3.5 (SD=.7)	3 (SD=.7)
The training program included what I would have wanted to learn as a new human supervisor of an autonomous driving system.	3.8 (SD=.8)	3.8 (SD=.9)	3.8 (SD=.8)
The training program was comprehensive, in that it covered enough of what I would need to know, and I'd be able to use an autonomous vehicle well-informed.	3.2 (SD=1)	3.3 (SD=1.2)	3.2 (SD=.9)
The training seems to have come from actual drivers as they experienced situations while supervising autonomous driving systems.	4.0 (SD=.7)	4.2 (SD=.7)	3.8 (SD=.7)
The training seems to have come from Tesla white papers and help manuals.	3.0 (SD=.9)	3.2 (SD=1)	2.9 (SD=.8)

Additionally, the average ratings were slightly higher for those participants that had above average test scores.

R3c2: Will the novices believe the training came from an authority in the domain (i.e., Tesla) or from peer end users (i.e., other Tesla FSD drivers)?

The participants were able to identify the source of the training accurately, as Tesla drivers, and not official documentation from Tesla. However, the practice and test stimuli came from driver videos posted on YouTube (YouTube, n.d.), which may have indicated to the participants that the training was crowdsourced from end users, and not official Tesla documentation.

9.8.2.4 Consumer Application: Oral Interview Questionnaire

R4a: What resources might a new Tesla FSD driver go to in order to find training material on the system?

Participants overwhelmingly stated that the first place they would go for training was the developer, Tesla. The second highest source was the Internet. Each participant listed an average of 6 sources (SD=2). For a detailed listing of responses, see Table D.1 in Appendix D.

R4b: With regards to the specific functionality the participants were trained on in the study, was the training sufficient in that that they would feel comfortable using that specific functionality of Tesla FSD? In what way(s)?

Participants felt the training was sufficient, and indicated it was more helpful than not. The reasons cited for the sufficiency of the training were that it was comprehensive and gave them a better overall understanding of the domain. A detailed listing of responses can be found in Table D.2, which is in Appendix D.

R4c: Does the training affect a potential Tesla FSD consumer's likelihood of purchasing the system? Does the awareness of failures identified in the training deter purchasers?

Participants were split on whether they'd be more or less likely to purchase a Tesla FSD vehicle after having completed the training. There were four participants each who stated that they would either be *more likely*, or maintained the *same likelihood*, and five participants who were *less likely*. Although participants found the training helpful, there appears to be some concern about the bugs and errors that were identified in the training. This demonstrates the aforementioned adverse effects of showing participants the failures of a system during training. A detailed listing of responses can be found in Table D.3, which is in Appendix D.

R4d: In what ways do participants understand autonomously driven vehicles now, after the training, that they did not previous to the training?

Participants generally reported a better understanding of autonomously driven vehicles after the training. However, they were surprised about the level of cautiousness taken by the system. For example, at a stop sign, if there are other vehicles in the intersection, the Tesla FSD will wait until all vehicles have cleared before proceeding, including those vehicles that arrived at the stop sign after the Tesla.

Participants also reported that they had a better understanding of the Tesla FSD's limitations. For a detailed listing of responses, see Table D.4, which is in Appendix D.

9.9 Discussion

This study evaluated the effectiveness of Explicit Rule Learning in the Tesla FSD domain. The results showed that participants' test scores were 30% higher in the Explicit Rule Learning (Experimental) condition versus those in the control (no training) condition, suggesting training with Explicit Rule Learning method was effective. This

indicates that Explicit Rule Learning can be a successful training method, providing learners with an understanding of the underlying principles of an intelligent software system, which led to more robust mental models, equipping the participants with the knowledge needed to understand and anticipate the decisions and output of these sophisticated systems. Using global, probabilistic, verbalizable if…then rules, complemented with factual and counterfactual exemplars and probabilities, participants had a better a overall understanding of a complex system after receiving the Explicit Rule Learning treatment.

Explicit Rule Learning was initially used as a successful training method for a simple machine learning image classifier and was easily adapted to a more sophisticated neural network AI system. The portability of the rule creation, Rule Card formation, and Practice with Feedback components confirmed the adaptability of this training method and its potential application to many types of AI/ML systems.

Additionally, it was possible to develop Explicit Rule Learning training material without access to an actual Tesla FSD vehicle. Resources were pooled from Tesla's website, peer-reviewed articles, special interest groups in social media and video sharing platforms, other online sources that provided information about the FSD system, and confirmed Tesla FSD drivers. Using this collective body of information, we were able to identify valid training content, cause-and-effect scenarios, and estimate probabilities that reflected the actual performance of these vehicles. We were able to identify principles which, when presented via Explicit Rule Learning, would help the learner gain a better

overall understanding of the system, and would provide the foundations needed to generalize this knowledge to future situations. . With due diligence, we were able to identify and validate the higher priority principles and events that occurred using the Tesla FSD autonomous vehicle system. With this tactic, we were able to focus on rules that described the more severe events (i.e., more dangerous or frequently occurring), and eliminate the more benign events that would not promote a global understanding of the underlying framework, logic and rationale of the system.

Showing the participants both the successes and failures of the system helped them to be better predictors of how the Tesla FSD system would react to certain situations. However, as expected, showing the failures did have an impact on their trust levels, with the trust decreasing as they learned more about the system. Previous research has found that declining trust can be recovered, with time, and instances of success (Sauer et al., 2016). One limitation of this study is that we were unable to study the trust level long term. If our study had results similar to Sauer's study, it's possible that with more time using the system, and with more examples of successes of the system, trust would have recovered to higher levels.

Another interesting finding was that novices (our participants did not have experience with autonomous vehicles) tend to seek training material from the developer, which was Tesla in this study. They also had consistent trust levels in the developer. This is inconsistent with the results from Mamun's (2023) study, whose participants were

experienced Tesla FSD drivers, and who tended to learn about the vehicle's functionality from other Tesla FSD drivers, or on their own.

This study provides data on empirically tested training content and presentation formats. In contrast, there are training resources, and even a service provider that offers Tesla FSD training, where the content was derived from personal and limited experiences of individual consumers or companies, presumably without the validation of sound research methods. While it is admittedly wise to act expediently, taking advantage of a budding industry, and take the opportunity to offer a new service to consumers, these forms of training do not have the benefit of fully testing, and empirically determining the most effective training content and methods.

With the burgeoning number of Tesla FSD beta testers in the wild, one doesn't have to look far to find reports documenting their experiences, with a focus on specific and local occurrences and decisions made by the autonomous vehicle system. Conversely, the aim of Explicit Rule Learning is to convey global principles to the learner, with the goal of a more robust mental model, and the ability to generalize the learned principles, replete with factual and counterfactual exemplars, to new, unseen circumstances. In this research, respondents were clear that Explicit Rule Learning provided a better overall understanding of the system, rather than outcomes specific to a defined set of circumstances.

The supplemental questionnaires provided additional information and investigatory support for this study. The results were used to get a clearer picture of the effect of this training on learners, and also, to identify possible factors that might contribute or counteract the effects of Explicit Rule Learning. For example, the results of the first study, using the MNIST classifier as stimuli, showed that counterfactual (negative) exemplars helped the participants more than factual (positive) exemplars. However, the final study, using the Tesla FSD autonomous vehicle system as stimuli, also provided counterfactual exemplars to the participants. The results of the final study, taken together with the Trust in Automation questionnaire results, as well as the Consumer Application Interview results indicate that there may be a need for trust recovery. Indeed, this requires a balance between the benefits of counterfactual exemplars and the regaining of trust in the AI/ML system. A future study might explore this further.

10 General Discussion

The studies described herein provide empirical evidence that Explicit Rule Learning accelerated the proficiency of learners of intelligent software systems. Participants had test scores up to 30% higher and reported a better overall understanding of the systems. The method was tested on a simple ML image classifier, and a more sophisticated neural network AI system. The Explicit Rule Learning method provided a more principled understanding, and a more robust mental model of these systems, as demonstrated by the participants' achievements as better predictors of how the system would respond to various scenarios, and the participants' descriptions of the systems' decision-making processes in varying scenarios.

10.1 Benefits of Explicit Rule Learning for AI/ML

Explicit Rule Learning is rooted in human cognition theories and contains global explanations of a system. The probabilistic, verbalizable if…then rules explain the logic, rationale, and categorization decisions made by the AI/ML systems, as well as boundary and contrastive category membership classifications.

The Rule Card component of Explicit Rule Learning is a one-page description of the rule, first described textually to the learner in an if…then statement. This description contains information on cases where the rule applies, and cases where the rule fails. Next, factual and counterfactual examples are presented to the learner as images or videos. These exemplars can be recalled as positive and negative depictions of the rule later on, when the learner is using the system and will be applying the learned rules. This is followed by

a visual representation of the probabilities of when the rule succeeds, and when it fails. Finally, a textual summary summarizes the rules probabilistic effectiveness.

The Rule Card is a well-rounded collection of training content, but there are additional layers of depth to this concise and simple rule presentation. First, the Rule Card contains system- or domain-specific nomenclature, key concepts, logic, and the processes that are used. Second, the learner can refer to the Rule Card in the future as a just-in-time resource, as needed. Third, is a depiction of a known problem space, structured in a way that will aid the representation of the problem in the learners' memory. Fourth, it reduces the cognitive load present when learning new material, providing a structured and standardized method of delivering rule-based training. Traditionally, exemplars burden the learner by putting the onus on them to make inferences. Lastly, the learner is guided through a combination of objects and features, signs, cues, and input that when taken as a whole make a difference in a system's output. Meaningful groups of factors in examples can benefit a learner's performance and problem-solving ability (Catrambone, 1996).

The Practice with Feedback component provides an active feedback loop, providing a safe and supportive environment for the learner to evaluate their proficiency, and adjust their behavior and strategies as needed. This visible progress increases the learners' motivation and engagement.

Explicit Rule Learning is not 100% predictive, but rather, it shortcuts the sensemaking process. It identifies decision-making rationale and priorities observed in an AI/ML system.

In summary, there are ways with which specific training methods can benefit learners: clear objectives (Moga and Cabaniss, 2014), relevance, engagement and interactivity (Chi and Wylie, 2014), structured and organized problem representation (Amarel, 1968), feedback, assessment, and measurable results (Watling and Ginsburg, 2019), practical application, flexibility, and adaptability. Here is a summary of these factors applied to Explicit Rule Learning:

- *Clear Objectives*: The rules describe causal events, signs and cues that make a difference, as well as a summary of the rule's effectiveness, and cases where it is not effective. These provide the learner with well-defined and specific principles of a system, which can be applied globally to future, novel situations.
- *Relevance*: The rules are selected such that they demonstrate global principles of the system and can be generalized to future situations. The rules illustrate the signs and cues that make a difference and contain factual and counterfactual exemplars.
- *Engaging and Interactive*: The Practice with Feedback component actively engages learners, making them more successful in forming a robust mental model of the system. The low-cost practice questions provide a no-pressure environment for the learner to have correct responses reinforced, and incorrect responses to be corrected. The Rule Card also helps with learning retention, providing memorable

and verbalizable rules, with visual exemplars, textual explanations, and probabilistic information.

- *Structured and Organized*: The Rule Cards are a structured and organized depiction of the rule. It is divided into logical sections, with a clear flow of content.
- *Feedback and Assessment*: The Practice with Feedback component is crucial to assess the learners progress. This helps the learner understand their strengths and identify areas that need improvement.
- *Practical Application*: The global approach allows the learner to transfer learned skills to new, unseen situations.
- *Measurable Results*: The learner receives immediate feedback in the Practice with Feedback component and develops a more robust mental model that can be consulted for new, unseen situations. The visibility of improvements made by the learner can be motivating.
- *Flexibility and Adaptability*: Explicit Rule Learning was adapted for two entirely different types of AI/ML systems (a simple ML image classifier and the more complicated Tesla FSD AI autonomous driving system). This demonstrates its ability to evolve and be incorporated into training for new technology and different types of AI/ML systems. The next section will detail this flexibility further.

10.2 Adaptability to Other AI/ML Systems

The adaptability of Explicit Rule Learning from a simple classifier to the more sophisticated real-world domain of a neural network AI system was accomplished with

modest effort. Even without access to the underlying algorithms, developers, white papers and source code, training material was developed within less than 6 weeks' time (from the point of initial exposure to the final development of training stimuli). There was no need for access to any proprietary information, and an actual Tesla FSD vehicle was not needed to identify the training content and create the Explicit Rule Learning material. This demonstrates the potential for using Explicit Rule Learning for many different types of intelligent software systems, including AI, and ML.

10.3 Expertise and Change Management

In order to get a clear understanding of the effect of the Explicit Rule Learning training intervention, participants were screened prior to completing the studies, and only novices in the domain proceeded with the study (e.g., participants in Study 5 had not previously driven an autonomous vehicle). In this way were able to attribute knowledge gains to the training methods (i.e., exemplars, Explicit Rule Learning for AI/ML). It is possible that participants who were experts in the domains would have produced different results in these studies. Training for experts might have a slightly different approach. For this reason, Explicit Rule Learning for AI/ML was conceptualized with flexibility and adaptability for varying levels of expertise.

Experts possess a deeper and more organized knowledge, are able to recognize and give meaning to patterns, and they solve problems differently from novices. As such, an expert's interpretation and application of a rule is different from that of a novice (Anderson et al., 1997). For example, in a study involving firefighters, Klein et al., (1986) found that robust mental models helped expert firefighters make intuitive

decisions based on pattern recognition and mental simulation rather than a systematic evaluation of the individual features and objects in the scenarios.

Ericsson and Kintsch (1995) explored a theory of long-term working memory in which experts store and manipulate complex information by leveraging their deep and organized knowledge of a domain. This allows an expert to be flexible with the application of a learned rule when presented with scenarios of varying context. Conversely, a novice might have a more rigid application of a rule, seeking exact matches to a rule's causal factors and expecting a singular outcome as defined by the rule. Therefore, the Explicit Rule Learning method will be slightly different when used to train experts in a domain.

10.3.1 Rule Content

First, the learning objectives selected as the content of the rule will differ. As described in Appendix F, the first step in developing the learning objectives that will be presented as rules is to perform a comprehensive data analysis of problems users encounter with the system. The learning objectives selected for novices would help them understand the basic underlying framework, logic, and rationale of the system. However, an expert already has that knowledge. Instead, the rules presented to experts in the domain would explain more anomalous occurrences or change management explanations. Experts, who have likely spent more hours interacting with the system are more likely to encounter scenarios that have a lower probability of occurring, but can nevertheless have a high impact. For example, in the autonomous vehicle domain, the Tesla FSD system may encounter road construction frequently, and the vehicle is able to adapt to narrowed and

temporary lane requirements and indicators. However, an expert (a human supervising the Tesla FSD system who drives considerably more hours than the average driver) may encounter the unique road construction configuration where there is only one lane shared by both directions of traffic, controlled with a temporary traffic light or humans who patrol the traffic, giving each direction its turn to traverse the one-lane road. In this example, the expert would be presented with a rule that would fine-tune their mental model.

The second type of rule an expert would be trained upon, a change management rule, might explain a difference in the system's operation due to an update, or some other change resulting from the dynamic nature of AI/ML systems. Changes in the human-automation teaming are ever-present. These changes might occur with the system or the user. For example, a system might be updated or improved, or there might be technological advances that require a modification to the system. A user will potentially achieve a higher level of expertise, or perhaps the users' goals evolve. In these cases, an expert can learn a rule that explains a newly enhanced or modified functionality of the system. For example, if an update to the system completely changes the way it processes input, a rule can guide the expert in the process of *unlearning* something. The rule card might explain the former and obsolete functionality and illustrate the new functionality.

10.3.2 Eliminate Practice with Feedback Component

The second difference in Explicit Rule Learning made to accommodate higher levels of expertise is that the expert will likely not need the Practice with Feedback component.

The deeper and more organized knowledge base of experts allows them to learn a rule, update their existing mental model, and proficiently and flexibly apply it to future scenarios without the need for practicing the learned material which would help encode the knowledge (Roediger and Karpicke, 2006).

Regardless of the reasons for updating Explicit Rule Learning Tutorials (e.g., modifications to the system, higher level of expertise in the users), updating the rules is a low-cost process. The existing data analysis (as described in Appendix F) can be updated to reflect the current state of the system. Learning objectives and exemplars can be revised and used to modify existing rules, or to replace obsolete rules.

10.4 Application to Real World AI/ML Intelligent Software Systems

The adaptability of Explicit Rule Learning, as seen by its effectiveness in a simple ML image classifier as well as a more sophisticated AI-based autonomous vehicle system, makes it useful as a training tool in today's world. The following describes a variety of ubiquitous AI/ML present in today's society, and examples of possible rules that might help accelerate the proficiency of learners.

Natural Language Processing Systems

Virtual Assistants: respond to user commands or questions, perform tasks, and provide information

Products: Amazon Alexa, Apple Siri

Sample Rules:

1. The system is designed to interpret keywords, such as commands ("open", "close"), questions ("where", "why"), intent ("buy", "look up"). If the keyword is used appropriately ("Open the garage door"), the system will respond as expected. Sometimes, the system will confuse a well-intended command ("Tell me where I can find *open* gym." "Where can I find an *open* road?"), and will respond incorrectly (Respond with gyms that are *open* or roads that are not *closed*.

2. The system has been trained in many facets, but sometimes it encounters limited flexibility in personalizing recommendations. For example, it may remember a person's preferences for a car rental (type of vehicle, pre-paid gas, etc.). However, if the same person has different preferences for a work-related vs a vacation-related car rental, it may make a mistake, and select the wrong preferences.

Natural Language Processing and Machine Learning Systems

Medical Diagnosis: analyze patient data, suggest treatment options, provide relevant research findings

Product: DeepMind

Sample Rules:

1. The system has been trained to make decisions based on the input it's been given, so it's important to give the system well-rounded and complete input to consider. If the system has received a comprehensive input (i.e., complete medical history, contextual information such as lifestyle factors or genetic predispositions, etc.), it will likely make an accurate diagnosis. However, if the system is given sparse input, it will make decisions based on limited and possibly erroneous assumptions.

2. The system is trained to be transparent in explaining how it arrived at its diagnosis. However, there are cases where a decision is derived after multiple layers of abstraction, and the explanation may not be structured in a way that a person readily understands. In these cases, the person should ask the system for the sub-components of its decisions, and analyze each sub-component before gaining an overall understanding of a diagnosis.

3. The system has been trained on data from limited demographics. If the input it is given was not in its training data, it may make errors in diagnoses.

Language Translation: transferring meaning from one language to another

Product: Google Translate

Sample Rules:

1. The system may make mistakes when translating colloquial or slang words and phrases, specific word pairings, cultural references, or domain-specific terms. When possible, these should be avoided or defined by the person, as otherwise, the system might provide an erroneous translation.

2. The system does not consider context or idioms. If context is relevant ("She *caught* a cold"), and the system isn't instructed to consider the context, it might make a mistake in the translation ("She *captured* a cold").

Natural Language Processing and Deep Learning Systems

Conversational AI: process, understand, and simulate human conversation

Product: ChatGPT (Chatbot)

Sample Rules:

1. When asking a complex or multi-part question, break it down before presenting them to the system. Otherwise, the responses may be erroneous, as the system may not understand the complete structure and hierarchy needed for an accurate response. The system will consider all sub-questions individually, without perceiving their role in the larger context.

2. The system has access to resources up until a specific date. If the question you are asking can apply to dates beyond the system's trained data, this limitation should

be taken into regard. When asking for date-specific information, make sure to inform the system of the dates you want it to consider.

Machine Learning Systems

Customer Relationship Management: maintain, analyze and harness business contact information

Product: Salesforce Einstein

Sample Rules:

1. The system makes sales-related predictions and recommendations based on patterns and trends in the entire global operation. If you want the system to consider a specific region, that should be indicated to the system.

2. When entering customer information, use the appropriate field on the form (order quantity in the quantity field, model number in the product model field, etc.) Otherwise the system will be inaccurate in its order history, projection, and/or customer history data.

3. The system may make a mistake in generating predictions and recommendations in specific contexts. For example, if a company has placed a large order recently, the system will give the order a specific weight, even if it's an anomaly and the only order the customer placed in the past ten years.

Content Recommendation: suggest personalized content based on user preferences and behavior

Product: Netflix Recommendation Engine

Sample Rules:

1. The system is trained on global data, which might place weight on movies that are popular, recently acclaimed, or recently trending programs. If your preference is two-fold ("movies of a specific genre" *and* "movies from the 20th century"), the system may make a poor recommendation by giving you movies of your preferred genre in a more recent year. If your preference is nuanced, the system may give you irrelevant recommendations.

2. If your preferences of programs shifts, the system may still consider your historical preferences when making recommendations. For example, if your history contains programs in the super-hero genre, but you've recently been watching documentaries, the system will make recommendations weighting the super-hero programs as this history is more dense under your profile.

Fraud Detection: analyze patterns and anomalies in transaction data to identify potential fraud cases

Product: PayPal Fraud Detection

Sample Rules:

1. Usually the system considers historic trends in a person's spending habits, but sometimes it will give a false positive erroneously. For example, if the system detects that two charges were made in two separate countries by one consumer in a 2-hour time period, it may flag the second transaction as a fraud. This false positive will not occur if the system considers the context (for example, one charge was in a restaurant in Sault Ste. Marie, MI, USA, and the second charge was at a gas station in nearby Sault Ste. Marie, Ontario, Canada).

2. The system may have false positives or miss cases of fraud due to the region and demographics it was trained on.

Image Organization: algorithmic analysis and recognition of visual input

Product: Google Photos

Sample Rules:

1. The system may not categorize images as expected. For example, although humans use shapes and forms, behavioral and sensory cues, and context to identify a dog, a system might use pixels and statistics, or rely on trained data.

Recognizing anomalies in system-based cues will help you to learn how it categorizes images.

2. The system may make a mistake by using similarity metrics. For example, a tiger might be classified as a housecat, due to their feature similarities, even though the tiger is usually not found in similar surroundings or environments.

Reinforcement Learning Systems

Game Playing: creating dynamic, personalized, adaptable experiences rooted in algorithms, that behave in a creative & intelligent way, mimicking human players

Product: DeepMind AlphaGo

Sample Rules:

1. The system usually perceives its position and state, and makes decision accordingly, but sometimes it has misperceptions and makes mistakes. For example, if the system hasn't been trained in specific situations, or if it perceives incomplete data or noise, it may overlook something such as a queen in position to capture its king in a chess match.

2. The system will learn an effective maneuver and use it repeatedly if it keeps winning. In these cases, you may be able to exploit the system's repetitive behavior by forming a strategy that surprises the system, causing it to lose as a result of its biased tendency of using a previously effective maneuver.

Computer Vision (image classification), Machine Learning

Facial Recognition: biometric mathematical mapping of an individual's facial features

Product: Facebook Facial Recognition

Sample Rules:

1. The system may have false negatives in some circumstances. For example, if the lighting is not good, if features are occluded or at variable angles, the system will not recognize the person. In contrast, sometimes the system will have false positives. For example, a nefarious actor might exploit a system with an attack ("presentation" or "spoofing") if they use a realistic photo of someone who has permissions to access a system.

2. The system may have been trained on specific regions or demographic groups. In these cases, the weights it places on specific features may differ from the people using the system.

For a high-level overview of the steps involved in identifying the learning objectives that form the content of the rules that will be presented on the Rule Card, please see Appendix F.

In summary, the adaptability and portability of Explicit Rule Learning makes it an effective method to train novices of intelligent software systems.

10.5 Limitations and Future Directions

In order to control and manipulate variables, the four studies described above used curated stimuli in this laboratory based research. The final study (Tesla FSD autonomous vehicle) contained real-world stimuli, captured in naturalistic environments; however, a true naturalistic environment might place the learners in the driver's seat of a Tesla in the real-world. Arguably, it would likely be sufficient to train Tesla FSD drivers online, as we did in this study; however, the testing scenarios could be more robust if tested in the wild, in an actual Tesla vehicle equipped with FSD functionality.

In this research, the participants were all novices in the domains. It is unknown how effective this training method would be for a more advanced user of an intelligent software system. A future study might measure the effectiveness of this method across various levels of users.

The participants in all of the studies were informed at the onset that they would be trained on a subset of the content upon which they would be tested. However, this was included after a verbose consent form, and before general instructions and expectations they should have for the subsequent study activities. Some participants reported that they were satisfied with their performance in the training portion, and felt surprised during the ensuing test, where they were tested on material in which they were not trained, despite the fact they were previously informed of this. This may account for the decrease in the Trust in Automation ratings. Some of the decrease might be attributed to the fact that the participants were shown the failures (counterfactual examples) of the systems; however,

the participants' reports of surprise due to their feeling of not being adequately prepared on untrained test items may have also contributed to the lower final Trust in Automation ratings.

Finally, in the future, Explicit Rule Learning will be researched on an even wider range of platforms, to test its effectiveness. Analysis will be done to determine Explicit Rule Learning's 1) adaptability to more types of AI/ML systems, 2) the potential of developing explicit, global, probabilistic, and verbalizable rules, and 3) the effectiveness of the training in a wider variety of AI/ML systems.

Lastly, a guide will be developed containing the procedure for developing Explicit Rule Learning for AI/ML. This will describe the steps of researching a system, performing data analysis, developing learning objectives, the construction of rules, Rule Card creation, and Practice with Feedback development.

Additionally, to the extent that Explicit Rule Learning could be applied to a more advanced AI system, it may be possible for XAI developers to incorporate this method into algorithmic or non-algorithmic XAI techniques when creating AI/ML explanations in the future. Many XAI methods have an economic basis as there is a trade-off between providing sufficient explanations to the end user, balancing the transparency and effort required by the system and the optimal level of information given to the end user that will result in their continued trust and use of the system. Explicit Rule Learning can be a low-cost training method for learners, supporting synergistic human-automation integration.

11 Reference List

Amarel, S. (1968). On representations of problems of reasoning about actions. In Michie (Ed.), *Machine Intelligence 3* (pp. 131-171). Evanston, IL: Edinburgh University Press.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological review*, *89*(4), 369.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of experimental psychology: learning, memory, and cognition*, *23*(4), 932.

APA Dictionary of Psychology. (n.d.). https://dictionary.apa.org/perception

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Atwood, J., Guo, F., Fitch, G., & Dingus, T. A. (2018). The driver-level crash risk associated with daily cellphone use and cellphone use while driving. *Accident Analysis & Prevention*, *119*, 149-154.

Baddeley, A., & Logie, R. (1999). Working Memory: The Multiple-Component Model. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 28-61). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139174909.005

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, 11(3), 211-227.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Bravo-Lillo, C., Cranor, L. F., Downs, J., & Komanduri, S. (2010). Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, *9*(2), 18-26.

Bruner, J. S. (1964). The course of cognitive growth. *American psychologist*, *19*(1), 1.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 1020.

Cattaneo, A. A., & Boldrini, E. (2017). Learning from errors in dual vocational education: Video-enhanced instructional strategies. *Journal of Workplace Learning*.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, *13*(2), 145-182.

Chi, M. T. H., Chiu, M. H. and Deleeuw, N. (1991). *Learning in a non-physical science domain: The human circulatory system*, Pittsburgh, PA: Learning Research and Development Center.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121-152.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219-243.

Chomsky, N. (1980). *Rules and Representations*. Columbia University Press, New York.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671-684.

Dikmen, M. & Burns, C.M. Autonomous driving in the real world: Experiences with Tesla Autopilot and summon. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Ann Arbor, MI, USA, 24–26 October 2016; pp. 225–228.

Dijksterhuis, A. (2004). Think different: the merits of unconscious thought in preference development and decision making. *Journal of personality and social psychology*, *87*(5), 586.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & Van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, *311*(5763), 1005-1007.

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary educational psychology*, *61*, 101859.

Edinger, A., & Goldstone, R. (2022). Getting Situated: Comparative Analysis of Language Models With Experimental Categorization Tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making*, 11(3), 225-238.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological review*, *102*(2), 211.

Fitts, P. M., & Posner, M. I. (1967). Human performance. brooks. *Cole, Belmont, CA*, *5*, 7-16.

Forbus, K. D., Hinrichs, E. T., Crouse, E. M., & Blass, J. (2020). Analogies versus Rules in Cognitive Architecture. *Proceedings of Advances in Cognitive Systems*.

Fried, L. S., & Holyoak, K. J. (1984). Induction of Category Distributions: A Framework for Classification Learning. *Learning, Memory*, 10(2), 234-257.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23(3), 183-209.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155-170.

Gentner, D. (1989). Analogical learning. Similarity and analogical reasoning, 199.

Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. MIT press.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In *Transfer of learning* (pp. 9-46). Academic Press.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology*, *278*(2), 563-577.

Gladwell, M. (2006). Blink: The power of thinking without thinking. adaptive unconscious: mental processes that work rapidly and automatically from relatively little information.

Goldstone, R. L., de Leeuw, J. R., & Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, *135*, 24-29.

Green, L. A., & Seifert, C. M. (2005). Translation of research into practice: why we can't "just do it". *The Journal of the American Board of Family Practice*, *18*(6), 541-545.

Grice HP. (1975). Logic and conversation. In *Syntax and Semantics*, ed. P Cole, J Morgan, pp. 41–58. New York: Academic.

Hampton, J. A. (1998). The Role of Similarity in How We Categorize The World. Holyoak, K., Gentner, D., and Kokinov, B., Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences, Bulgaria, 19-30.

Hase, P., & Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?. *arXiv preprint arXiv:2005.01831*.

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., & Zuckerman, O. (2019, May). Can children understand machine learning concepts? The effect of uncovering black boxes. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-11). Data labeling and evaluation help children's mental model of ML system.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68-73.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Holyoak, K. J., & Morrison, R. G. (Eds.). (2005). *The Cambridge handbook of thinking and reasoning* (Vol. 137). Cambridge: Cambridge University Press.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology*, *101*(1), 116.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418.

James, W. (1890). *The principles of psychology* (Harvard ed., 2 vols.). New York: Holt.

Johnson-Laird, P. N. (1989). Mental models. In M. I. Posner (Ed.), Foundations of cognitive science (pp. 469–499). The MIT Press.

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690*.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063, pp. 218-226). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4), 237. State that our intuitive predictions are based on representativeness of past outcomes. Both in how close it matches, and if we apply the right category/representation.

Kahneman, D., & Tversky, A. (1982). The Psychology of Preferences. *Scientific American*, *246*(1), 160–173. http://www.jstor.org/stable/24966506

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of educational psychology*, *93*(3), 579.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive science*, *8*(3), 255-273.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, *29*.

Klausmeier, H. J., & Hooper, F. H. (1974). Conceptual Development and Instruction. *Review of Research in Education*, *2*, 3–54. https://doi.org/10.2307/1167158

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. *Decision making in action: Models and methods*, *5*(4), 138-147.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986, September). Rapid decision making on the fire ground. In *Proceedings of the human factors society annual meeting* (Vol. 30, No. 6, pp. 576-580). Sage CA: Los Angeles, CA: Sage Publications.

Klein, G. (2018). Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, *33*(2), 83-88.

Klein, G., Hintze, N., & Saab, D. (2013, May). Thinking inside the box: The ShadowBox method for cognitive skill development. In *Proceedings of the 11th International Conference on Naturalistic Decision Making, Marseille, France* (Vol. 24, pp. 121-124).

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data–frame theory of sensemaking. In *Expertise out of context* (pp. 118-160). Psychology Press.

Körber, M. (2018, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association* (pp. 13-30). Springer, Cham.

Krause, J., Perer, A., & Bertini, E. (2018, August). A user study on the effect of aggregating explanations for interpreting machine learning models. In *ACM KDD Workshop on Interactive Data Exploration and Analytics*.

Kuhl, U., Artelt, A., & Hammer, B. (2023). Let's go to the Alien Zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *Frontiers in Computer Science*, *5*, 20.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, August). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675-1684).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Levine, G. M., Halberstadt, J. B., & Goldstone, R. L. (1996). Reasoning and the weighting of attributes in attitude judgments. *Journal of personality and social Psychology*, *70*(2), 230.

Linja, A.; Mamun, T.I.; Mueller, S.T. When Self-Driving Fails: Evaluating Social Media Posts Regarding Problems and Misconceptions about Tesla's FSD Mode. *Multimodal Technol. Interact.* 2022, *6*, 86. https://doi.org/10.3390/mti6100086

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural processes*, *66*(3), 309-332.

Mamun, T.I. (2023). *Investigating Collaborative Explanation as an Explainable AI (XAI) Method in Autonomous Driving.* [Unpublished manuscript].

Medin, D. L., & Heit, E. (1999). Categorization. In *Cognitive science* (pp. 99-143). Academic Press.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual review of psychology*, 35(1), 113-138.

Merlhiot, G., & Bueno, M. (2022). How drowsiness and distraction can interfere with take-over performance: A systematic and meta-analysis review. *Accident Analysis & Prevention*, *170*, 106536.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Moga, D. E., & Cabaniss, D. L. (2014). Learning objectives for supervision: Benefits for candidates and beyond. *Psychoanalytic Inquiry*, *34*(6), 528-537.

Mozannar, H., Satyanarayan, A., & Sontag, D. (2022, June). Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 5, pp. 5323-5331).

Mueller, S. T. (2020). Cognitive anthropomorphism of AI: How humans and computers classify images. *Ergonomics in Design,* 28(3), 12-19.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. (Defense Advanced Research Projects Agency XAI Program). https://arxiv.org/pdf/1902.01876.

Mueller, S. T., & Klein, G. (2011). Improving users' mental models of intelligent software tools. *IEEE Intelligent Systems*, 26(2), 77-83. Experiential User Guide to help develop more accurate mental models.

Mueller, S. T., Tan, S. Y. Y., Linja, A., Klein, G., & Hoffman, R. R. (2021). *Authoring guide for Cognitive Tutorials for Artificial Intelligence: Purposes and the Methods Development*. DARPA XAI Technical Report.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic bulletin & review*, *15*, 465-494.

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of experimental psychology: Learning, memory, and cognition*, 11(1), 70.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, 92(3), 289.

Neubauer, C., Matthews, G., & Saxby, D. (2012, September). The effects of cell phone use and automation on driver performance and subjective state in simulated driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 1987-1991). Sage CA: Los Angeles, CA: Sage Publications.

Nickerson, R. S., & Adams, M. J. (1979). Long-term memory for a common object. *Cognitive psychology*, 11(3), 287-307.

Norman, D. A. (1988). *The psychology of everyday things*. Basic books.

Nosofsky, R. M., & Little, D. R. (2010). Classification response times in probabilistic rule-based category structures: Contrasting exemplar-retrieval and decision-boundary models. *Memory & Cognition*, *38*(7), 916-927.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, 50(2), 530-556.

Nushi, G. B. B., Kamar, E., Lasecki, W. S., & Horvitz, D. S. W. E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance.

Ohlsson, S. (1996). Learning from performance errors. *Psychological review*, 103(2), 241.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, 97(2), 185.

Paul, E. J., & Ashby, F. G. (2013). A neurocomputational theory of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience*, *7*, 177.

Polk, T. A., The Great Courses (Firm), & Kanopy (Firm). (2018). *The Learning Brain*. Teaching Company, LLC. https://books.google.com/books?id=OEcbtgEACAAJ

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-52).

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE transactions on systems, man, and cybernetics*, (3), 257-266.

Rau, M. A., Aleven, V., & Rummel, N. (2010, June). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In *International conference on intelligent tutoring systems* (pp. 413-422). Springer, Berlin, Heidelberg.

Roediger, H. L. & Goff, L. M. (1999) Chapter 17: Memory. In *A Companion to Cognitive Science,* Bechtel, W. & Graham, G. (eds) Blackwell, Malden MA.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Rosch, E. (1978). Principles of categorization. *Cognition and categorization*, 27-48.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3), 382-439.

Rutledge-Taylor, M., Lebiere, C., Thomson, R., Staszewski, J., & Anderson, J. R. (2012). A comparison of rule-based versus exemplar-based categorization using the ACT-R architecture. In *Proceedings of the 21st Conference on Behavior Representation in Modeling and Simulation, BRIMS Society: Amelia Island, FL*.

SAE International. (2021).SAE Levels of Driving Automation Refined for Clarity and International Audience. SAE Levels of Driving Automation Refined for Clarity and International Audience. Retrieved May22, 2022.

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767-780.

Smith, E. E. (1995). Concepts and categorization. In E. E. Smith & D. Osherson (Eds.), *Invitation to cognitive science: Vol. 3. Thinking* (2nd ed., pp. 3-33). Cambridge, MA: MIT Press.

Smith, E. E. (2008). The case for implicit category learning. *Cognitive, Affective, & Behavioral Neuroscience*, 8(1), 3-16.

Smith, E. & Medin, D. (1981/1999). The exemplar view (ch. 3 of their *Categories and Concepts*). In Margolis and Laurence (1999).

Smith, E. E., & Medin, D. L. (2002). The exemplar view. *Foundations of cognitive psychology: Core readings*, 277-292.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*(2-3), 167-196.

Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory & Cognition*, 22(4), 377-386.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of learning and memory*, *82*(3), 171-177.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder?. *Cognitive psychology*, 23(3), 457-482.

Thorndike, E. L. (1908). Memory for paired associates. *Psychological review*, *15*(2), 122.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science (New York, NY)*, *185*(4157), 1124-1131.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, *46*(4), 197-221.

van der Meij, H., & Flacke, M. L. (2020). A review on error-inclusive approaches to software documentation and training. *Technical communication, 67*(1), 83-95.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404.

van Gog, T., Rummel, N., & Renkl, A. (2019). Learning how to solve problems by studying examples. In J. Dunlosky & K. Rawson (Eds.), The Cambridge Handbook of Cognition and Education (pp. 183-208). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235631.009

van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review*, *17*, 147-177.

Vosniadou, A. & Ortony, S. (editors). *Similarity and Analogical Reasoning*. Cambridge, UK: Cambridge Univ. Press, 1989.

Wang, C., Cavanagh, A. J., Bauer, M., Reeves, P. M., Gill, J. C., Chen, X., Hanauer, D.I., & Graham, M. J. (2021). A framework of college student buy-in to evidence-based teaching practices in STEM: the roles of trust and growth mindset. *CBE—Life Sciences Education*, *20*(4), ar54.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016, March). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 109-116). IEEE.

Watling, C. J., & Ginsburg, S. (2019). Assessment, feedback and the alchemy of learning. *Medical education*, *53*(1), 76-85.

Watson, J. B. (1925). *Behaviorism*. New York: W. W. Norton.

Wick, M. R., & Thompson, W. B. (1992). Reconstructive expert system explanation. *Artificial Intelligence*, *54*(1-2), 33-70.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: introspection can reduce the quality of preferences and decisions. *Journal of personality and social psychology*, *60*(2), 181.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*, 202-238.

YouTube. (n.d.). Online video sharing and social media platform. https://www.youtube.com.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017, March). User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 307-317).

Zhang, J. (2019). Cognitive functions of the brain: perception, attention and memory. *arXiv preprint arXiv:1907.02863*.

Zhao, H., Ma, J., Zhang, Y., & Chang, R. (2022). Mental workload accumulation effect of mobile phone distraction in L2 autopilot mode. *Scientific reports*, *12*(1), 16856.

A Rule Cards

MNIST Studies 1-4

Figure A.1

Rule Card: MNIST Study, 3-2 Rule 1

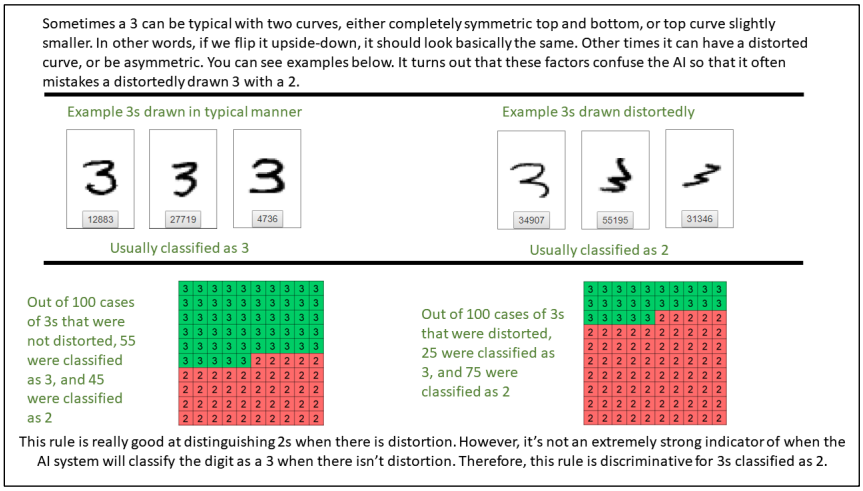


Figure A.2

Rule Card: MNIST Study, 3-2 Rule 2

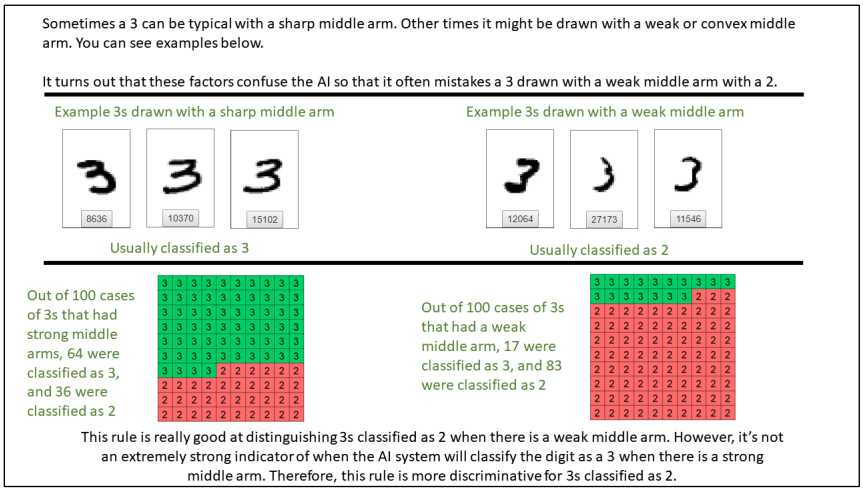


Figure A.3

Rule Card: MNIST Study, 1-5 Rule 1

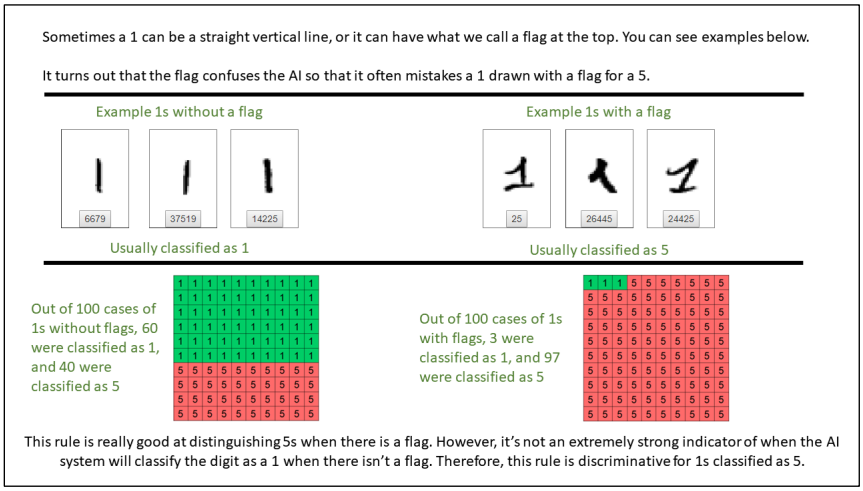


Figure A.4

Rule Card: MNIST Study, 1-5 Rule 2

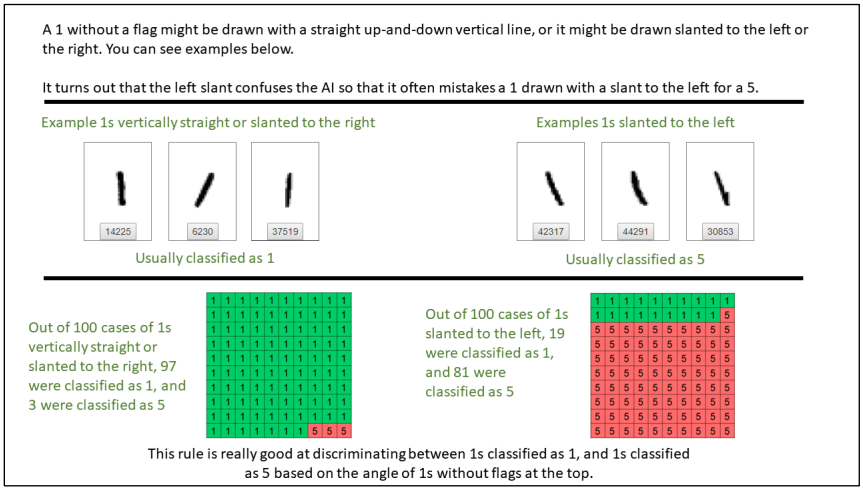


Figure A.5

Rule Card: MNIST Study, 0-6 Rule 1

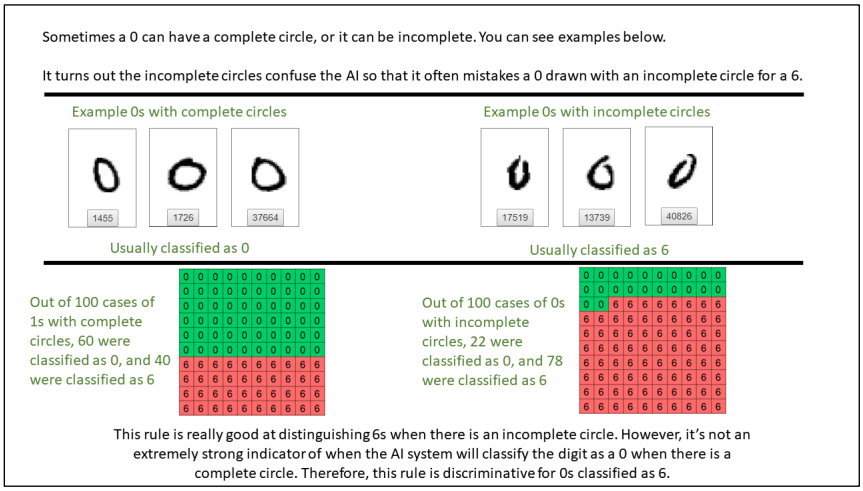


Figure A.6

Rule Card: MNIST Study, 0-6 Rule 2

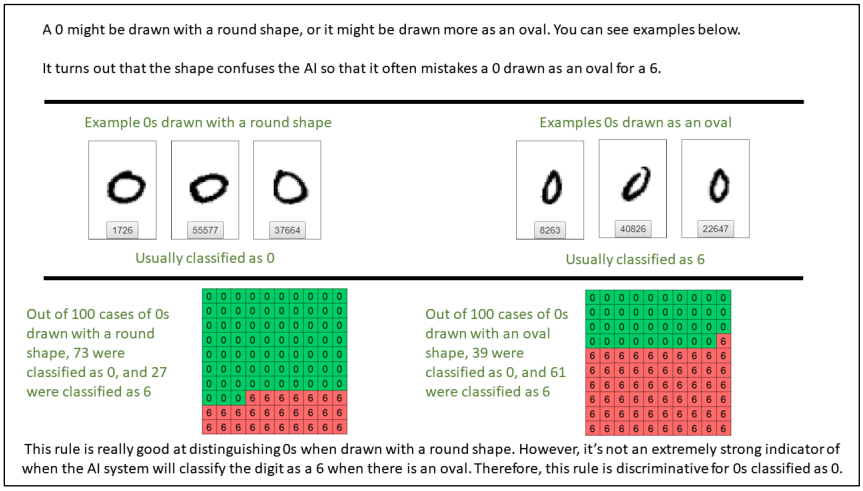


Figure A.7

Rule Card: MNIST Study, 4-9 Rule 1

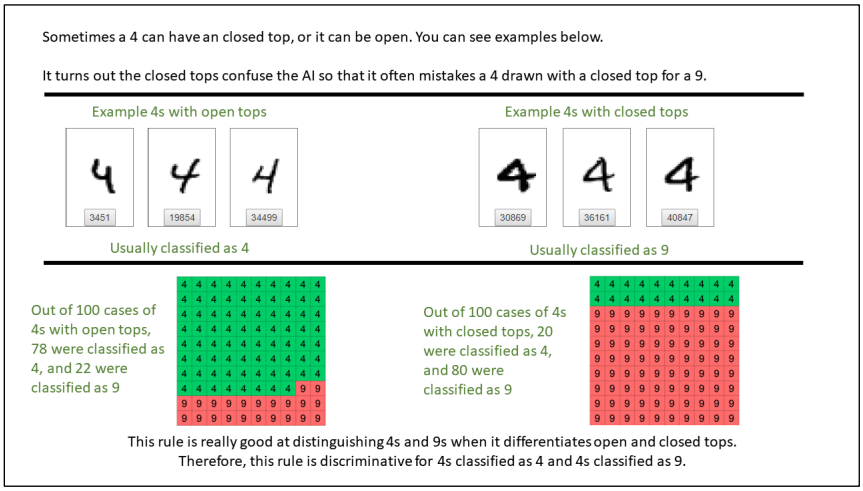


Figure A.8

Rule Card: MNIST Study, 4-9 Rule 2

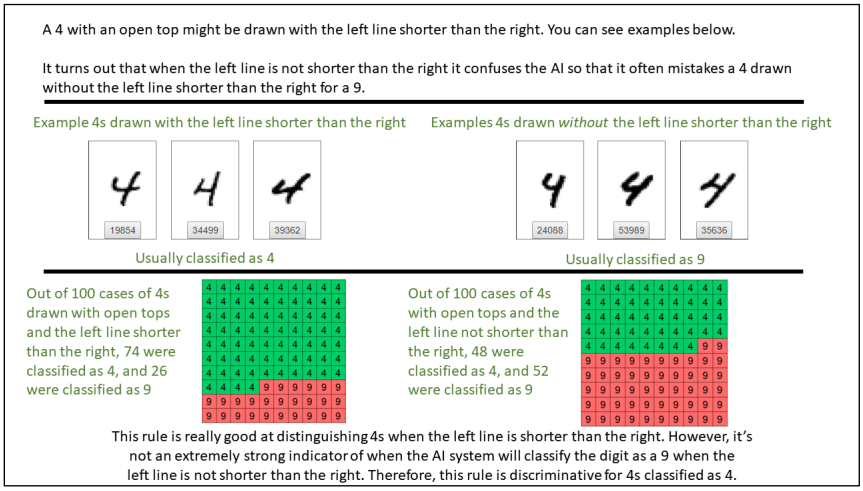


Figure A.9

Rule Card: Tesla FSD Study, Lane Rule 1

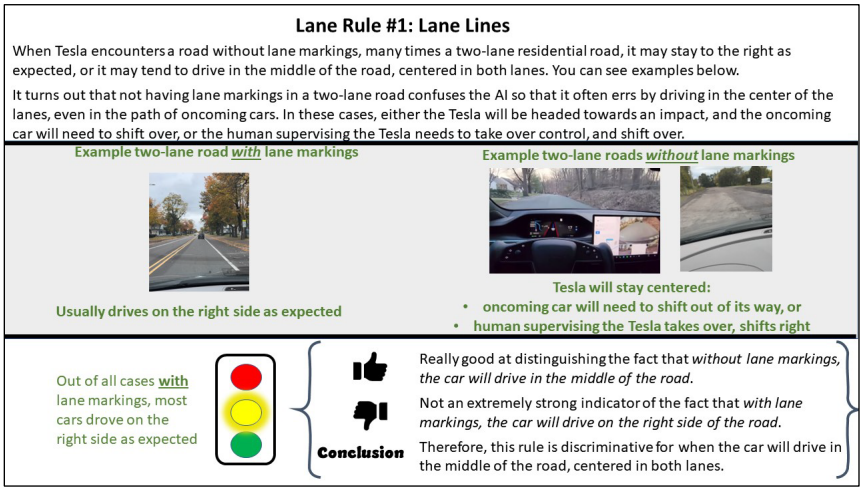


Figure A.10

Rule Card: Tesla FSD Study, Lane Rule 2

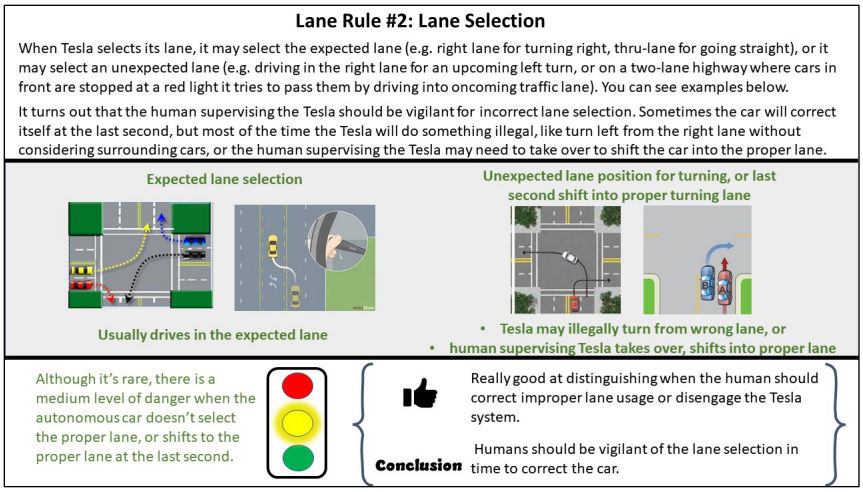


Figure A.11

Rule Card: Tesla FSD Study, Timid Approach Rule 1

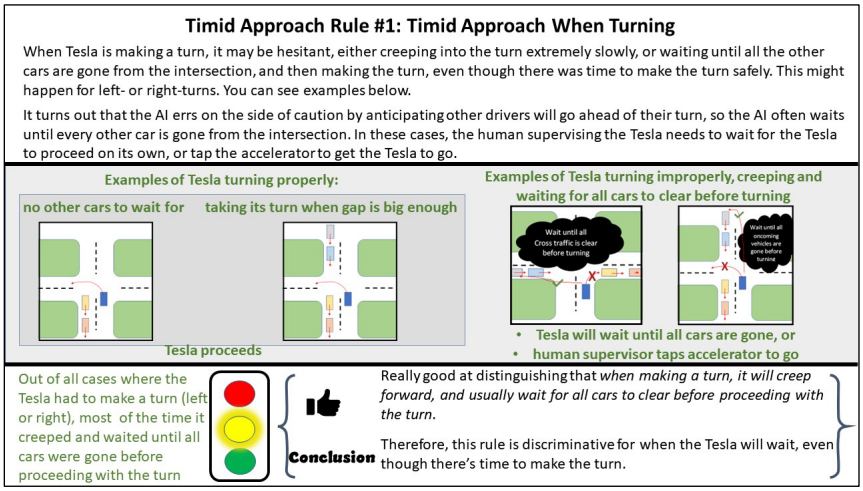
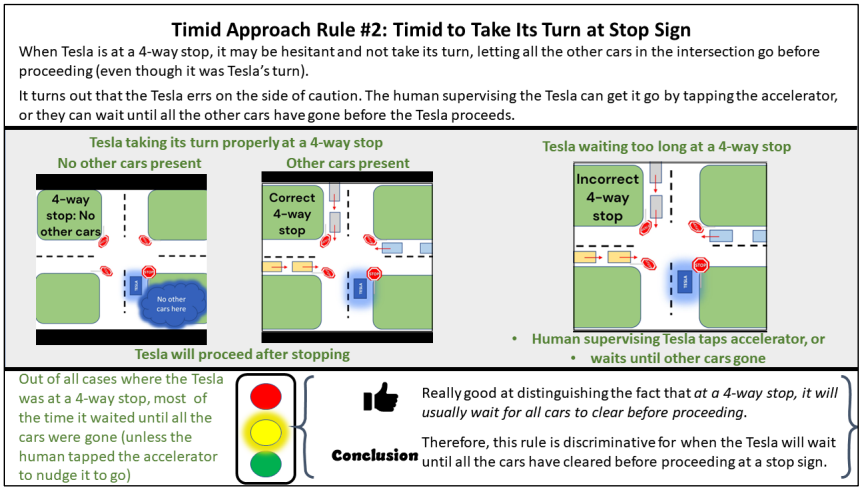


Figure A.12

Rule Card: Tesla FSD Study, Timid Approach Rule 2



B Demographics Questionnaire

1. Gender
    a. Male
    b. Female
    c. Non-binary
    d. Prefer not to say
2. Age
3. How old were you when you obtained a full U.S. driver's license (not a permit, intermediate or restricted license)?
4. Approximately how far (in miles) have you driven during the last 12 months?
5. Do you have a smartphone? If yes, please list the 2 most common ways that you use it while you are driving.

Table B.1

Summarized Responses: Cellphone Usage While Driving with Counts

Usage	**Count**
Music/Podcast	39
Maps/Navigation	30
Phone call	16
Other	3
Texting	2
Do not use	1

C User Experience Questionnaire

(Adapted from a parallel study in the Veinott/Mueller Lab, researcher: TI Mamun)

Note: The response options (5-item Likert scale), as shown in statement 1 below, were the same for all 5 statements.

1. I was able to identify the autonomous vehicle's subsequent actions for each video.

   [] Strongly agree

   [] Agree

   [] Neither agree nor disagree

   [] Disagree

   [] Strongly disagree

2. The training program included what I would have wanted to learn as a new human supervisor of an autonomous driving system.
3. The training program was comprehensive, in that it covered enough of what I would need to know, and I'd be able to use an autonomous vehicle well-informed.
4. The training seems to have come from actual drivers as they experienced situations while supervising autonomous driving systems.
5. The training seems to have come from Tesla white papers and help manuals.

D Trust in Automation Questionnaire

Figure D.1

Körber Trust in Automation Questionnaire

		Strongly disagree	Rather disagree	Neither disagree nor agree	Rather agree	Strongly agree	No response
1	The system is capable of interpreting situations correctly.	①	②	③	④	⑤	○
2	The system state was always clear to me.	①	②	③	④	⑤	○
3	I already know similar systems.	①	②	③	④	⑤	○
4	The developers are trustworthy.	①	②	③	④	⑤	○
5	One should be careful with unfamiliar automated systems.	①	②	③	④	⑤	○
6	The system works reliably.	①	②	③	④	⑤	○
7	The system reacts unpredictably.	①	②	③	④	⑤	○
8	The developers take my well-being seriously.	①	②	③	④	⑤	○
9	I trust the system.	①	②	③	④	⑤	○
10	A system malfunction is likely.	①	②	③	④	⑤	○
11	I was able to understand why things happened.	①	②	③	④	⑤	○
12	I rather trust a system than I mistrust it.	①	②	③	④	⑤	○
13	The system is capable of taking over complicated tasks.	①	②	③	④	⑤	○
14	I can rely on the system.	①	②	③	④	⑤	○
15	The system might make sporadic errors.	①	②	③	④	⑤	○
16	It's difficult to identify what the system will do next.	①	②	③	④	⑤	○
17	I have already used similar systems.	①	②	③	④	⑤	○
18	Automated systems generally work well.	①	②	③	④	⑤	○
19	I am confident about the system's capabilities.	①	②	③	④	⑤	○

Table D.1

Trust Factors: Mean Responses Pre-Study, Post-Training, and Post-Test

Trust Factor	Pre-study	Post-training	Post-test	ANOVA (Type II Wald χ2 tests)
Familiarity	2.77	2.36	2.22	**χ^2 (27.9) = 2, p < .05**
Intention of Developers	3.60	3.53	3.47	χ^2 (2.1) = 2, p = 0.36
Propensity to Trust	2.77	2.59	2.47	χ^2 (4.8) = 2, p = 0.09
Reliability/ Competence	3.10	2.73	2.55	**χ^2 (65.1) = 2, p < .05**
Understanding/ Predictability	3.11	3.23	2.99	**χ^2 (7.2) = 2, p < .05**
Trust in Automation	3.19	2.81	2.63	**χ^2 (33.9) = 2, p < .05**

Note. Bolded ANOVA results are statistically significant.

E Consumer Application: Oral Interview Questionnaire

"As you saw in the training, the Tesla Full-Self Driving vehicle operates autonomously, with the human in the driver's seat, supervising and ready to disengage the system, or take over as needed.

Imagine you are in the market to purchase a Tesla, with the Full-Self Driving technology enabled."

1. Where are some places you'd find training on the FSD (AV) system?

Table E.1

Q1 Responses: Training Sources with Counts

Source	Count
Tesla-provided documentation (manual, website, Tesla dealer)	30
Online Resource (non-conversational; e.g., search engine result, YouTube video)	18
Existing Tesla drivers	10
Autodidactic/Self-teach/Self-experience	7
Online Resource (conversational; e.g., threaded social media, special interest group blog)	3
Other	2

2. Would any of the training you just completed help you learn enough use the vehicle? In what way?

Table E.2

Q2 Responses: Helpfulness of Training, Reasons, Counts

Was the training helpful?	***Why/Why Not?***	***Count***
Yes (9)	Sufficient Training/Comprehensive	11
	Better Overall Understanding of Autonomous Vehicle Domain	6
	Awareness of Indicative Cues/Signs/Signals	1
No (3)	Insufficient Training/Need More Training	10
	Automation Failures Caused Confusion	6
	Recap of Training	7

3. If you were considering purchasing an autonomously-driven vehicle before taking this training, would you be more or less likely to do so now, after the training?

Table E.3

Q3 Responses: Likelihood of Purchasing an Autonomous Vehicle, Reasons, Counts

Likelihood	Reason for likelihood	Count
Less Likely	It has bugs and errors	9
	Doesn't have enough capabilities	3
	Less trust after training	1
	Need for constant vigilance	1
	Wouldn't use it all the time	1
More Likely	New awareness	3
	Would be good for long road trips	3
	Have more confidence now	2
	Never had interest in getting one before	1
	Nice to have	1
	Saves energy and money	1

4. In what ways do you understand autonomously driven vehicles now, after the training, that you didn't before?

Table E.4

Q4 Responses: Understanding of Autonomous Vehicles After Training with Counts

New Understanding	Count
Better understanding of its functionality	12
Tesla's cautiousness/safety first (overprotective, more caution than previously thought)	10
Better understanding of limitations/not as effective as human manually driving	8
Technology is still being developed/room for improvement	5
Similar to human in its thinking/decisions	1
Making great progress with AVs	1
Already knew about AVs/didn't learn anything new	1

F Selection of Learning Objectives for Rule Cards

The following is a high-level overview of the steps involved in identifying and selecting the learning objectives that will subsequently be turned into Explicit Rule Learning training. The goal is that the collection of rules will give the learner a comprehensive understanding of the underlying framework, logic, rationale, and goals of the system. This global understanding of the system should enable the learner to understand and predict the system's output and generalize the knowledge to future unseen scenarios. The rules should educate the user on how the system works in the domain rather than describe the underlying algorithms of the system.

The steps listed below are based on the foundations set forth by the Authoring guide for Cognitive Tutorials for Artificial Intelligence: Purposes and the Methods Development (Mueller et al., 2021).

1. A systematic analysis of the system is performed to learn about the system, its potential users, and the goals of the users of the system. Additionally, gaps, mistakes, workarounds, errors, tendencies of the system, and unexpected results are collected.
    a. The output of this step is:
        i. a list of the problems identified, and
        ii. vignettes, stories, and examples that will be used as factual and counterfactual exemplars.
    b. The goal is to identify possible learning objectives. However, although efforts should be made to collect a comprehensive and representative list,

at this stage the efforts should be focused on collecting data, and not on developing the specific learning objectives or rules, or prioritizing the items.

c. The sources for this data collection might come from a wide array of resources, including (but not limited to):

- Interviews of expert users, programmers, developers, and designers of the system,
- Academic/proprietary institutions,
- White papers,
- Observation/shadowing of users interacting with the system,
- Online courses/tutorials,
- Online forums/communities,
- YouTube/videos,
- FAQs/bug databases,
- Conferences/workshops,
- Sandbox/hands-on/manually interacting with the system, and/or
- Similar tools/systems.

2. Prioritization and concretization of the learning objectives.

a. The final learning objectives are identified after the data collection, with consideration to:

- Learners,
- How the system will be used,
- Other available training,

- Resources available for creation and implementation of the training,
- Time,
- Practicality,
- Importance of learning objectives (i.e., severity of outcome, such as a higher level of danger or illegality, etc.)
- Instructing the learner on global principles using local cases as exemplars, and/or
- The goal of a generalizable understanding.

b. The learning objective should result in rules that express:
    - The most likely output given a set of input,
    - Any assumptions or constraints,
    - Feature salience,
    - Feature weights,
    - Possible reasoning,
    - Any boundary conditions,
    - Modifications that change the output,
    - Possible logic, and
    - The context in which the rule applies.

c. The learning objective should result in a rule that is:
    - Clearly stated,
    - Verbalizable,
    - Probabilistically true, and

- Global.

Here are sample rule cards for a Virtual Assistant Natural Language Processing and Facial Recognition AI systems:

Figure F.1

Rule Card: Virtual Assistant

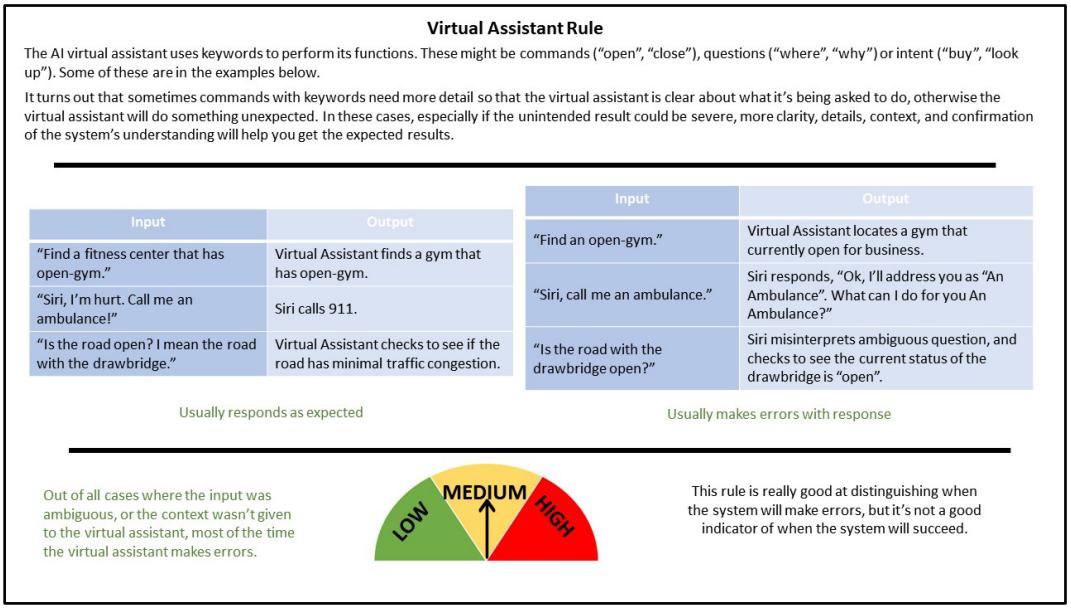
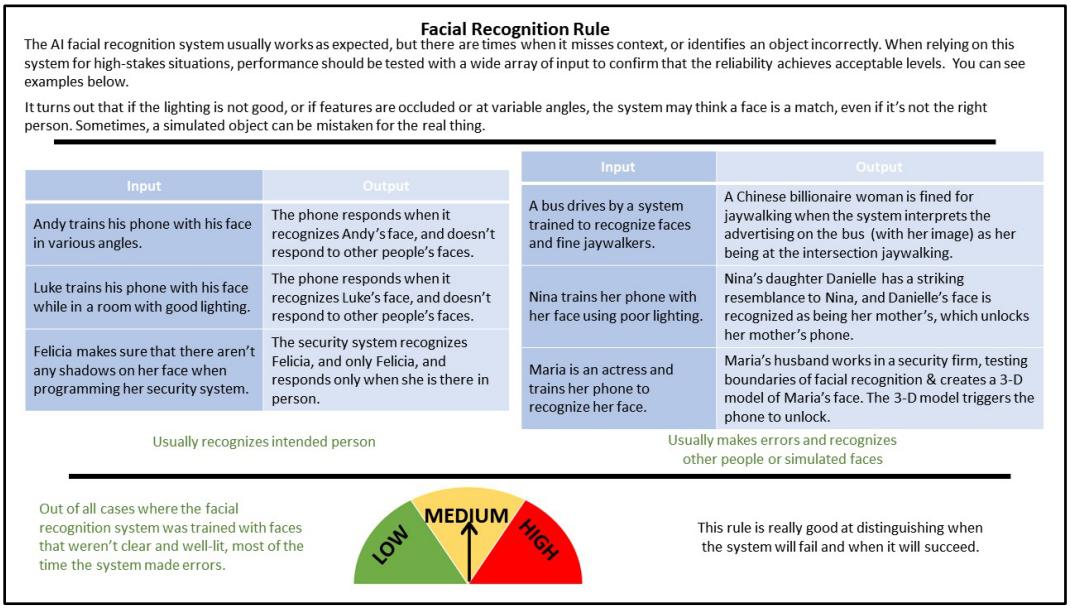


Figure F.2

Rule Card: Facial Recognition



Facial Recognition Rule

The AI facial recognition system usually works as expected, but there are times when it misses context, or identifies an object incorrectly. When relying on this system for high-stakes situations, performance should be tested with a wide array of input to confirm that the reliability achieves acceptable levels. You can see examples below.

It turns out that if the lighting is not good, or if features are occluded or at variable angles, the system may think a face is a match, even if it's not the right person. Sometimes, a simulated object can be mistaken for the real thing.

Input	Output
Andy trains his phone with his face in various angles.	The phone responds when it recognizes Andy's face, and doesn't respond to other people's faces.
Luke trains his phone with his face while in a room with good lighting.	The phone responds when it recognizes Luke's face, and doesn't respond to other people's faces.
Felicia makes sure that there aren't any shadows on her face when programming her security system.	The security system recognizes Felicia, and only Felicia, and responds only when she is there in person.

Usually recognizes intended person

Input	Output
A bus drives by a system trained to recognize faces and fine jaywalkers.	A Chinese billionaire woman is fined for jaywalking when the system interprets the advertising on the bus (with her image) as her being at the intersection jaywalking.
Nina trains her phone with her face using poor lighting.	Nina's daughter Danielle has a striking resemblance to Nina, and Danielle's face is recognized as being her mother's, which unlocks her mother's phone.
Maria is an actress and trains her phone to recognize her face.	Maria's husband works in a security firm, testing boundaries of facial recognition & creates a 3-D model of Maria's face. The 3-D model triggers the phone to unlock.

Usually makes errors and recognizes other people or simulated faces

Out of all cases where the facial recognition system was trained with faces that weren't clear and well-lit, most of the time the system made errors.

This rule is really good at distinguishing when the system will fail and when it will succeed.

ɔs://www.bbc.com/news/technology-46357004